Surface Structure Predictions and Development of Global Exploration Tools

DISSERTATION

zur Erlangung des akademischen Grades

Doctor rerum naturalium (Dr. rer. nat.)

im Fach Chemie

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät

der Humboldt-Universität zu Berlin

von

Herrn Magister Radosław Stanisław Włodarczyk

Präsident der Universität: Prof. Dr. Jan-Hendrik Olbertz

Dekan der Fakultät: Prof. Dr. Elmar Kulke

Gutachter/in: 1. Prof. Dr. Dr. h.c. Joachim Sauer

2. Prof. Dr. Marek Sierka

3. Prof. Dr. Bernd Hartke

Tag der mündlichen Prüfung: 06. Mai 2015

Abstract

This work is a contribution in the field of theoretical chemistry and surface science. The joint computational and experimental studies investigated the atomic structure of ultrathin silica and iron-doped silica films formed on the Ru(0001) surface and water films formed on the MgO(001) surface. The atomic structure models were obtained using either the educated guess approach or the genetic algorithm that was designed and implemented within the DoDo package. The properties simulated for the resulting models are in a very good agreement with the experimental data (scanning tunnelling microscopy, infrared spectroscopy). The successful structure determination using the DoDo program shows that the genetic algorithm technique is capable of systematic and extensive exploration of the energy landscapes for 2D-periodic systems.

Keywords:

global exploration methods, genetic algorithms, periodic boundary conditions, density functional theory, ultrathin films, surface science, oxides

Zusammenfassung

Diese Arbeit ist ein Beitrag zur theoretischen Chemie sowie zur Oberflächenchemie. Durch Kombination von computergestützten und experimentellen Untersuchungen wird die atomare Struktur von dünnen SiO_2-Filmen auf Ru(0001)-Unterlagen, von eisendotierten SiO_2-Filmen auf diesen Unterlagen und von H_2O-Filmen auf MgO(001)-Oberflächen bestimmt. Die atomaren Strukturmodelle wurden entweder mit dem neu entworfenen und im Paket DoDo implementierten genetischen Algorithmus oder mittels auf Sachkenntnis gestützter Vermutungen erhalten. Die simulierten Eigenschaften der so erhaltenen Strukturen stimmen sehr gut mit den experimentellen Daten (Raster-Tunnel-Mikroskopie, Infrarot-Spektroskopie) überein. Die erfolgreiche Strukturbestimmung mithilfe des DoDo-Programms zeigt, dass genetische Algorithmen zur systematischen und extensiven Erkundung der Energielandschaften 2D-periodischer Systeme geeignet sind.

Schlagwörter:

Globale Optimierungsmethoden, genetischer Algorithmus, periodische Randbedingungen, Dichtefunktionaltheorie, dünne Filme, Oberflächenchemie, Oxide

Contents

1. Introduction 1

1.1 Outline of the Dissertation 1

1.2 List of co-authored references 2

2. The potential energy surface paradigm 4

2.1 Concept of the potential energy surface 4

2.2 Energy landscapes 8

2.3 Local exploration of energy landscapes 10

2.3.1 Structure optimization 10

2.3.2 Transition structure search 11

2.4 Global exploration of energy landscapes 14

2.4.1 Educated guess approach 14

2.4.2 Molecular dynamics method 15

2.4.3 Metropolis Monte Carlo technique 16

2.4.4 Genetic algorithms 17

2.4.5 Simulated annealing scheme 27

2.4.6 Basin hopping scheme 28

2.4.7 Particle swarm optimization technique 29

2.4.8 Metadynamics method 30

2.4.9 Umbrella sampling scheme 32

3. Educated guess used for atomic structure determination 34

3.1 Introduction 34

3.2 Crystalline silica sheet grown on the Ru(0001) surface 35

3.3 Electronic structure of ultrathin crystalline silica films 36

3.4 Monolayer, bilayer, and three-dimensional networks 39

3.5 Atomic structure of ultrathin Fe-silicate film 41

4. The DoDo genetic algorithm package 48

4.1 General package description 49

4.2 DoDo implementation 50

4.2.1 Initial structure models 50

4.2.2 Local optimization 54

4.2.3 Survival of the fittest 55

4.2.4 Crossover 55

4.2.5 Similarity recognition 58

4.2.6 Dynamic steady-state algorithm 62

5. Application of the DoDo genetic algorithm 66

5.1 Introduction 66

5.2 Case study: Ordered water monolayer on the MgO(001) surface 68

5.2.1 Exploration of energy landscapes 69

5.2.2 Refinement of results 73

5.2.3 Comparison of simulated and experimental properties 75

5.3 Conclusions 76

6. Summary 78

Appendix: keywords used in the DoDo input file 80

List of abbrevations 86

References 88

Declaration 92

Complete list of the candidate's publications 94

Curriculum Vitae 96

Acknowledgement 98

1. Introduction

1.1 Outline of the Dissertation

Understanding of the structure-property relationship is essential for effective tailoring of materials with various physical or chemical properties. Diverse model systems such as single crystal surfaces, thin films or gas phase clusters are experimentally prepared and analyzed in order to gain more insight into this relation. However, even for such “simplified” systems, characterization of the atomic structure can be a challenging task. In principle, the structure of these nanoscale materials might resemble the respective bulk structures. However, the arrangement of atoms or even the stoichiometric proportions can be unusual or even counterintuitive as well. Recently, many joint experimental and theoretical studies provided detailed and reliable information about the atomic and electronic structure of model systems. These successful studies were based on high-quality experimental data that complemented the computational effort. In this Dissertation, I present selected investigations in which I was responsible for the computational part.

In Chapter 2, I introduce selected theoretical concepts and techniques for structure determination. First, I evoke the paradigm of the potential energy surface (PES). Then, selected methods for local exploration of the PES are illustrated. Finally, chosen techniques for global exploration of the PES (with particular emphasis on the genetic algorithms) are described.

The third Chapter focuses on examples of successful application of the educated guess approach, which is the method of the first choice for atomic structure determination problems. These investigations are selected joint theoretical and experimental projects. The studies are given in a chronological order. They demonstrate how an initially small project evolved into a comprehensive investigation. It can be also seen how theoretical effort became increasingly more important for a study that was at first driven experimentally.

In the fourth Chapter, I describe the DoDo genetic algorithm package. It is a tool programmed in order to perform an automated systematic search in the conformational space. Thus, it represents an alternative to the intuition-guided educated guess approach acknowledged in

Chapter 3. The implementation of the most important genetic operators is described in detail. The list of keywords used to control the program is presented in the Appendix.

In the fifth Chapter, I present an example of successful application of the DoDo genetic algorithm package to the atomic structure determination problem. I provide details of the investigation of the ordered water monolayer structures on the MgO(001) surface and I demonstrate the course of the genetic algorithm run for this case.

Below, I provide a list of publications that serve as the basis for this Dissertation.

1.2 List of co-authored references

W1 C. Büchner, L. Lichtenstein, X. Yu, J. A. Boscoboinik, B. Yang, W. E. Kaden, M. Heyde, S. K. Shaikhutdinov, **R. Włodarczyk**, M. Sierka, J. Sauer, and H.-J. Freund, "Ultrathin Silica Films: The Atomic Structure of Two-Dimensional Crystals and Glasses," *Chem. Eur. J.* **20**, 9176 (2014).

W2 D. Löffler, J. J. Uhlrich, M. Baron, B. Yang, X. Yu, L. Lichtenstein, L. Heinke, C. Büchner, M. Heyde, S. Shaikhutdinov, H. J. Freund, **R. Włodarczyk**, M. Sierka, and J. Sauer, "Growth and Structure of Crystalline Silica Sheet on Ru(0001)," *Phys. Rev. Lett.* **105**, 146104 (2010).

W3 **R. Włodarczyk**, M. Sierka, J. Sauer, D. Löffler, J. J. Uhlrich, X. Yu, B. Yang, I. M. N. Groot, S. Shaikhutdinov, and H. J. Freund, "Tuning the Electronic Structure of Ultrathin Crystalline Silica Films on Ru(0001)," *Phys. Rev. B* **85**, 085403 (2012).

W4 B. Yang, W. E. Kaden, X. Yu, J. A. Boscoboinik, Y. Martynova, L. Lichtenstein, M. Heyde, M. Sterrer, **R. Włodarczyk**, M. Sierka, J. Sauer, S. Shaikhutdinov, and H.-J. Freund, "Thin Silica Films on Ru(0001): Monolayer, Bilayer and Three-Dimensional Networks of [SiO_4] Tetrahedra," *Phys. Chem. Chem. Phys.* **14**, 11344 (2012).

W5 L. Lichtenstein, C. Büchner, B. Yang, S. Shaikhutdinov, M. Heyde, M. Sierka, **R. Włodarczyk**, J. Sauer, and H.-J. Freund, "The Atomic Structure of a Metal-Supported Vitreous Thin Silica Film," *Angew. Chem., Int. Ed.* **51**, 404 (2012).

W6 J. A. Boscoboinik, X. Yu, B. Yang, F. D. Fischer, **R. Włodarczyk**, M. Sierka, S. Shaikhutdinov, J. Sauer, and H.-J. Freund, "Modeling Zeolites with Metal-Supported Two-Dimensional Aluminosilicate Films," *Angew. Chem., Int. Ed.* **51**, 6005 (2012).

W7 **R. Włodarczyk**, J. Sauer, X. Yu, J. A. Boscoboinik, B. Yang, S. Shaikhutdinov, and H.-J. Freund, "Atomic Structure of an Ultrathin Fe-Silicate Film Grown on a Metal: A Monolayer of Clay?," *J. Am. Chem. Soc.* **135**, 19222 (2013).

W8 **R. Włodarczyk**, M. Sierka, K. Kwapień, J. Sauer, E. Carrasco, A. Aumer, J. F. Gomes, M. Sterrer, and H.-J. Freund, "Structures of the Ordered Water Monolayer on MgO(001)," *J. Phys. Chem. C* **115**, 6764 (2011).

2. The potential energy surface paradigm

In the first part of this chapter, the separation of electronic and nuclear wavefunction, *i.e*. the Born-Oppenheimer approximation, is introduced. This approximation applied to the time-independent Schrödinger equation leads to definition of the potential energy surface. The potential energy surface (PES) is a function which for an investigated system describes the relationship between atomic positions and the energy for the given electronic state. In the second part, selected properties of the energy landscapes are briefly described. Finally, the most significant methods used for local and global exploration of the energy landscape are outlined. Introductory Sections 2.1 and 2.2 are based on the textbooks by D. Wales[1] and I. N. Levine.[2] Sections 2.3 and 2.4 describing local and global PES exploration techniques are based on the paper by H. B. Schlegel.[3] Method-specific sources are given in the text. The function used throughout this Chapter as an example of the PES is taken from Ref. 4.

2.1 Concept of the potential energy surface

In the time-independent Schrödinger equation

$$\hat{H}\Psi = E\Psi, \tag{2.1.1}$$

$\hat{H}$ is the Hamilton operator that acts on the wavefunction Ψ, which represents a stationary state, yielding the energy E of that state. In atomic units, the non-relativistic Schrödinger equation (2.1.1) for system consisting of nuclei and electrons treated as point masses takes the form:

$$\left[\hat{T}_{\mathrm{n}} + \hat{T}_{\mathrm{e}} + \hat{V}\right]\Psi(\boldsymbol{x},\boldsymbol{X}) = E\Psi(\boldsymbol{x},\boldsymbol{X}), \tag{2.1.2}$$

where $\hat{T}_{\mathrm{n}}$ is the nuclear kinetic energy operator, $\hat{T}_{\mathrm{e}}$ is the electronic kinetic energy operator, $\hat{V}$ is the potential energy operator, $\boldsymbol{x}$ and $\boldsymbol{X}$ stand for electronic and nuclear coordinates, respectively. All these terms are explicitly given below:

$$\hat{T}_{\mathrm{n}} = -\frac{1}{2}\sum_{\alpha}\frac{1}{M_{\alpha}}\nabla_{\alpha}^{2}, \tag{2.1.3a}$$

$$\hat{T}_{\mathrm{e}} = -\frac{1}{2}\sum_{i}\nabla_i^2, \tag{2.1.3b}$$

$$\hat{V} = \hat{V}_{\mathrm{nn}} + \hat{V}_{\mathrm{ne}} + \hat{V}_{\mathrm{ee}} = \sum_{\alpha}\sum_{\beta>\alpha}\frac{Z_\alpha Z_\beta}{r_{\alpha\beta}} - \sum_{\alpha}\sum_{i}\frac{Z_\alpha}{r_{\alpha i}} + \sum_{i}\sum_{j>i}\frac{1}{r_{ij}}. \tag{2.1.3c}$$

In Eqns. (2.1.3a-c), α and β refer to nuclei, i and j to electrons, M_α denotes the mass of the nucleus α, Z_α the charge of the nucleus α, and $r_{\alpha\beta}$ the distance between indicated particles. The potential energy terms in Eq. 2.1.3c describe nuclei-nuclei, electron-nuclei, and electron-electron Coulomb interactions, respectively.

Hamiltonian given in Eq. 2.1.2 is a function of nuclear and electronic spatial coordinates treated on an equal footing. Solving this problem becomes cumbersome even for small systems. However, since the rest mass of an electron is ~10^3 times smaller than the rest mass of a proton, the nuclear configuration ($\boldsymbol{X}$) can be considered as much more inert than the electron distribution ($\boldsymbol{x}$). Thus, the electrons should instantaneously adjust to changes in the positions of nuclei. Under such assumptions, coupling between electronic and nuclear degrees of freedom can be replaced by parametrical dependence of the mobile electron density on the inert nuclear positions. Then, the overall wavefunction of the system can be defined as:

$$\Psi(\boldsymbol{x},\boldsymbol{X}) = \Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}), \tag{2.1.4}$$

where $\Psi_{\mathrm{n}}(\boldsymbol{X})$ denotes the nuclear wavefunction and $\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})$ stands for the electronic wavefunction. Substituting Eq. 2.1.4 into Eq. 2.1.2 gives:

$$\left[\hat{T}_{\mathrm{n}} + \hat{T}_{\mathrm{e}} + \hat{V}\right]\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}) = \tag{2.1.5}$$

$$= \underbrace{\hat{T}_{\mathrm{n}}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X})}_{\boxed{1}} + \underbrace{\hat{T}_{\mathrm{e}}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X})}_{\boxed{2}} + \underbrace{\hat{V}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X})}_{\boxed{3}} =$$

$$= E\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}).$$

Due to the separation of variables, $\nabla_\alpha\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X}) = 0$ and $\nabla_i\Psi_{\mathrm{n}}(\boldsymbol{X}) = 0$. Thus, the kinetic energy terms $\boxed{1}$ and $\boxed{2}$ can be simplified as follows:

$$\boxed{1}: \hat{T}_{\mathrm{n}}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}) = \tag{2.1.6a}$$

$$= -\frac{1}{2}\sum_{\alpha}\frac{1}{M_{\alpha}}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\nabla_{\alpha}^{2}\Psi_{\mathrm{n}}(\boldsymbol{X}) = T_{\mathrm{n}}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}),$$

$$\boxed{2}: \hat{T}_{\mathrm{e}}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}) = \tag{2.1.6b}$$

$$= -\frac{1}{2}\sum_{i}\Psi_{\mathrm{n}}(\boldsymbol{X})\nabla_{i}^{2}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X}) \;=\; T_{\mathrm{e}}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}),$$

and the potential energy term $\boxed{3}$:

$$\boxed{3}: \hat{V}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}) = \tag{2.1.6c}$$

$$= \hat{V}_{\mathrm{nn}}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}) + \hat{V}_{\mathrm{ne}}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}) + \hat{V}_{\mathrm{ee}}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X})$$

$$= V_{\mathrm{nn}}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}) + V_{\mathrm{ne}}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}) + V_{\mathrm{ee}}\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}).$$

The T_{n} and T_{e} eigenvalues introduced in Eqns. (2.1.6a-b) denote the kinetic energy of nuclei and the kinetic energy of electrons, respectively. The V_{nn}, V_{ne}, and V_{ee} eigenvalues used in Eq. 2.1.6c denote the potential energy of nuclei-nuclei, electron-nuclei, and electron-electron Coulomb interactions, respectively.

Due to the separation of variables, $\hat{T}_{\mathrm{n}}$ acts solely on $\Psi_{\mathrm{n}}(\boldsymbol{X})$ and $\hat{T}_{\mathrm{e}}$ solely on $\Psi_{\mathrm{e}}(\boldsymbol{x};\boldsymbol{X})$, *i.e.* the explicit coupling between $\boldsymbol{x}$ and $\boldsymbol{X}$ is neglected in the kinetic energy terms. The explicit coupling between electrons and nuclei is conserved only in the $\hat{V}_{\mathrm{ne}}$ term. Such a formulation is known as the Born-Oppenheimer approximation.

Within the Born-Oppenheimer approximation, the electrons move in a static electric field generated by nuclei according to the $\hat{V}_{\mathrm{ne}}$ operator, are interacting with each other in a way described by the $\hat{V}_{\mathrm{ee}}$ operator, and are characterized by the kinetic energy determined by the $\hat{T}_{\mathrm{e}}$ operator. Since only these three terms depend explicitly on the electronic coordinates, they are grouped in the so-called electronic Hamiltonian $\hat{H}_{\mathrm{e}}$:

$$\hat{H}\Psi_{\mathrm{e}}(\boldsymbol{x};\,\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}) = \tag{2.1.7}$$

$$= \left[\hat{T}_{\mathrm{n}} + \hat{V}_{\mathrm{nn}} + \underbrace{\hat{V}_{\mathrm{ne}} + \hat{V}_{\mathrm{ee}} + \hat{T}_{\mathrm{e}}}_{\hat{H}_{\mathrm{e}}}\right] \Psi_{\mathrm{e}}(\boldsymbol{x};\,\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X})$$

$$= \left[\hat{T}_{\mathrm{n}} + \hat{V}_{\mathrm{nn}} + \hat{H}_{\mathrm{e}}\right] \Psi_{\mathrm{e}}(\boldsymbol{x};\,\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}).$$

The electronic wavefunction $\Psi_e(\boldsymbol{x};\,\boldsymbol{X})$ introduced in Eq. 2.1.4 is the solution of the electronic Hamiltonian problem:

$$\hat{H}_{\mathrm{e}}\Psi_{\mathrm{e}}(\boldsymbol{x};\,\boldsymbol{X}) = E_{\mathrm{e}}\Psi_{\mathrm{e}}(\boldsymbol{x};\,\boldsymbol{X}). \tag{2.1.8}$$

The eigenvalue E_{e} is called the electronic energy.

Since the electric potential of nuclei defined by the $\hat{V}_{\mathrm{ne}}$ operator depends on the atomic coordinates $\boldsymbol{X}$ (*cf*. Eq. 2.1.3c), the whole electronic Hamiltonian $\hat{H}_{\mathrm{e}}$ is a function of $\boldsymbol{X}$ as well:

$$\hat{H}_{\mathrm{e}}(\boldsymbol{X})\,\Psi_{\mathrm{e}}(\boldsymbol{x};\,\boldsymbol{X}) = E_{\mathrm{e}}(\boldsymbol{X})\,\Psi_{\mathrm{e}}(\boldsymbol{x};\,\boldsymbol{X}). \tag{2.1.9}$$

Solving this equation for a particular set of nuclear coordinates $\boldsymbol{X_i}$ yields electronic wavefunction $\Psi_{\mathrm{e}}(\boldsymbol{x};\,\boldsymbol{X_i})$ and electronic energy $E_{\mathrm{e}}(\boldsymbol{X_i})$ for the given conformation. The remaining potential energy term, $\hat{V}_{\mathrm{nn}}$, describes nuclei-nuclei repulsion. This entity depends only on nuclear coordinates and is constant for the given $\boldsymbol{X_i}$:

$$\left[\hat{H}_{\mathrm{e}}(\boldsymbol{X}) + \hat{V}_{\mathrm{nn}}(\boldsymbol{X})\right] \Psi_{\mathrm{e}}(\boldsymbol{x};\,\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}) = \tag{2.1.10}$$

$$= \left[E_{\mathrm{e}}(\boldsymbol{X}) + V_{\mathrm{nn}}(\boldsymbol{X})\right] \Psi_{\mathrm{e}}(\boldsymbol{x};\,\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X})$$

$$= V(\boldsymbol{X})\,\Psi_{\mathrm{e}}(\boldsymbol{x};\,\boldsymbol{X})\Psi_{\mathrm{n}}(\boldsymbol{X}).$$

The function $V(\boldsymbol{X})$ is called the potential energy surface (PES). The value $V(\boldsymbol{X_i})$ is the potential energy of a system in the given conformation $\boldsymbol{X_i}$.

2.2 Energy landscapes

Potential energy surface $V(\boldsymbol{X})$ defined in Eq. 2.1.10 represents the potential energy of the given system as a function of nuclear coordinates. $V(\boldsymbol{X})$ can be expanded in a Taylor series centered at $\boldsymbol{X_0}$:

$$V(\boldsymbol{X}) = V(\boldsymbol{X_0}) + \nabla V(\boldsymbol{X_0})^{\mathrm{T}}(\boldsymbol{X} - \boldsymbol{X_0}) + \frac{1}{2}(\boldsymbol{X} - \boldsymbol{X_0})^{\mathrm{T}}\nabla^2 V(\boldsymbol{X_0})(\boldsymbol{X} - \boldsymbol{X_0}) + \cdots \quad (2.2.1)$$

where $\nabla V(\boldsymbol{X_0})$ is the gradient vector and $\nabla^2 V(\boldsymbol{X_0})$ is the Hessian matrix, both evaluated at $\boldsymbol{X} = \boldsymbol{X_0}$. If $\nabla V(\boldsymbol{X_0}) = 0$, then $\boldsymbol{X_0}$ represents a stationary point and Eq. 2.2.1 is reduced to:

$$V(\boldsymbol{X}) = V(\boldsymbol{X_0}) + \frac{1}{2}(\boldsymbol{X} - \boldsymbol{X_0})^{\mathrm{T}}\nabla^2 V(\boldsymbol{X_0})(\boldsymbol{X} - \boldsymbol{X_0}) + \cdots \quad (2.2.2)$$

Equation (2.2.2) can be further simplified by applying the harmonic approximation, *i.e.* neglecting the terms of higher than a second order:

$$V(\boldsymbol{X}) = V(\boldsymbol{X_0}) + \frac{1}{2}(\boldsymbol{X} - \boldsymbol{X_0})^{\mathrm{T}}\nabla^2 V(\boldsymbol{X_0})(\boldsymbol{X} - \boldsymbol{X_0}). \quad (2.2.3)$$

Properties of a stationary point $\boldsymbol{X_0}$ depend on the curvature of the PES at this point. Within the harmonic approximation, the curvature of the PES at the stationary point $\boldsymbol{X_0}$ is determined by the eigenvalues of the Hessian matrix $\nabla^2 V(\boldsymbol{X_0})$ calculated at this point. Each positive eigenvalue corresponds to one eigenvector for which an infinitesimal displacement increases the potential energy of the system while each negative eigenvalue corresponds to one eigenvector for which an infinitesimal displacement decreases the potential energy of the system. Thus, the stationary points with only positive eigenvalues are local minima while simultaneous presence of negative and positive eigenvalues characterises saddle points. The number of negative eigenvalues defines the order of the saddle point, *i.e.* the number of mutually orthogonal displacement vectors that can be followed to decrease the potential energy of the system. The PES can be divided into basins of attraction of local minima (Figure 2.1). A point on the PES belongs to the basin of attraction of the given local minimum if a local optimization procedure started from that point leads to that given minimum. The PES landscape (*energy landscape*) can become complicated even for relatively small systems since the number of local minima grows exponentially with the system size.[5]

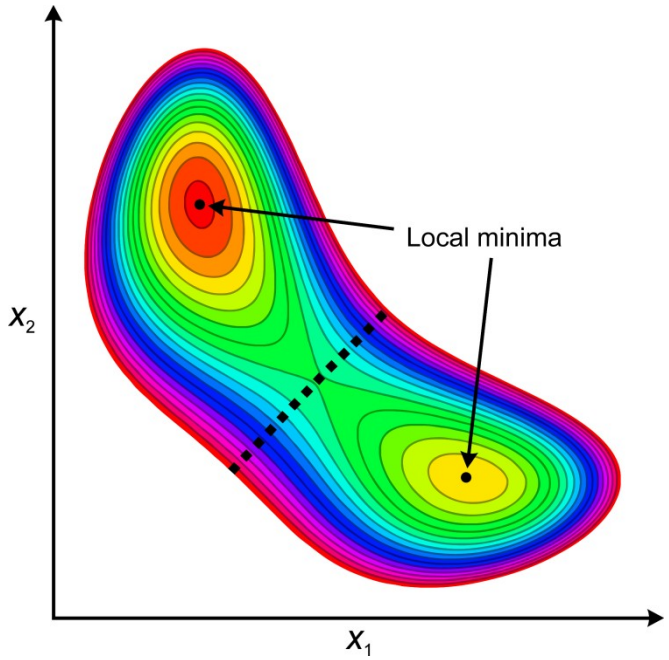

Figure 2.1: An example of a two-dimensional potential energy surface. The dashed line represents the boundary between the basins of attraction of two indicated minima.

Many properties, *e.g.* structure, dynamics, thermodynamics, or reactivity of a system, can be derived from knowledge of the PES. Thus, the exploration of the energy landscape is essential to understand or to predict many chemical and physical properties of a system. For example: (1) local minima on the PES represent (meta)stable configurations; (2) the deepest minimum, *i.e.* the global minimum, corresponds to the thermodynamically most stable system configuration at the temperature of 0K; (3) the saddle points correspond to transition structures; (4) shape of the vicinity of a minimum determines the vibrational spectrum of the corresponding structure. Exploration of energy landscapes can be performed locally or globally and the decision which technique should be applied depends on the investigated property.

2.3 Local exploration of energy landscapes

Local exploration techniques focus on gathering information on a local structure of the PES, *e.g.* related to the given local minimum, saddle point, or the pathway connecting two nearby minima. In this Section, selected local exploration techniques are briefly described.

2.3.1 Structure optimization

Structure optimization is a process during which the initial spatial arrangement of atoms is iteratively adjusted in the way that minimizes the potential energy of the system. Here, conceptually the simplest approach, *i.e.* the steepest descent method, is briefly described.

If $V(\boldsymbol{X_n})$ is the potential energy at the point $\boldsymbol{X_n}$ and $\nabla V(\boldsymbol{X_n})$ is the gradient of the potential energy at this point, then the optimization step in the steepest descent method proceeds in the direction of the negative gradient of the potential energy and is given as:

$$\boldsymbol{X_{n+1}} = \boldsymbol{X_n} - \alpha \nabla V(\boldsymbol{X_n}), \tag{2.3.1}$$

with α being a scaling factor. In case of infinitesimally small α, the optimization follows the path of the *steepest descent* of the energy. For the practical applications, α is system-dependent and can also vary during the optimization procedure. An example of optimization path obtained by the steepest descent method is shown in Figure 2.2.

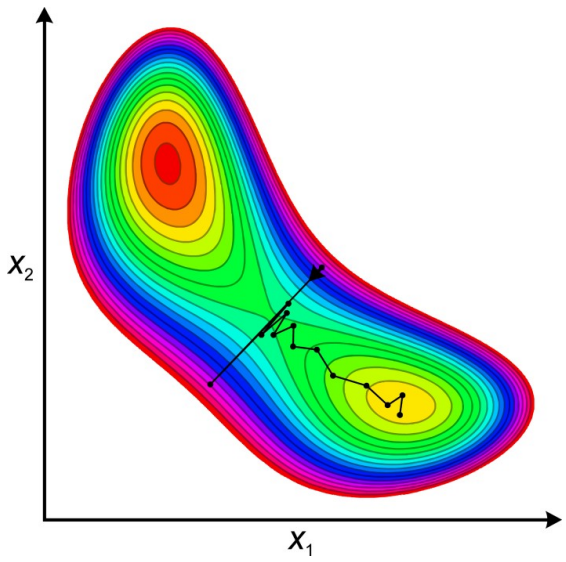

Figure 2.2: An example of an optimization path obtained by the steepest descent method on a two-dimensional potential energy surface.

The steepest descent method is based only on the knowledge of the local slope. The information obtained in the previous step is discarded immediately after each optimization step. It is well known that for functions with narrow valleys, *i.e.* fast change of gradient in one direction and almost flat profile in other directions, the optimization procedure can be drastically hindered by oscillations (zigzag path). A solution to this problem is to introduce the memory of the previous steps to the working equations, as it is done in the conjugated gradient method.

The algorithms discussed so far are based solely on the slope (gradient) of the PES, *i.e.* they utilize only the linear term of the local Taylor expansion of the PES (*cf.* Eq. 2.2.1). The methods that include the quadratic term of this expansion are called the Newton-type algorithms. The Newton's method requires computation of the full matrix of the second derivatives (Hessian matrix) $H(\boldsymbol{X_0})$ and then its inverse $H^{-1}(\boldsymbol{X_0})$. Modern algorithms omit these computationally demanding steps and apply an estimate of the inverted Hessian matrix $B^{-1}(\boldsymbol{X_0})$ instead of the exact $H^{-1}(\boldsymbol{X_0})$. The $B^{-1}(\boldsymbol{X_0})$ elements are updated on-the-fly from the gradients calculated in each step. Methods that are based on the approximated inverse of Hessian matrix are called the quasi-Newton methods with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm being the most prominent representative of this group of algorithms.

2.3.2 Transition structure search

Chemical reactions can be described by reaction mechanisms, *i.e.* by a collection of elementary steps that describe the progress of the reaction in time. The structures present in the given reaction mechanism can be reflected on the potential energy surface and connected by paths. An example of PES for a reaction for which reactants and products occupy adjacent basins of attraction and the energy profile associated with the selected reaction path is depicted in Figure 2.3. The point $\boldsymbol{X_s}$ indicated in Figure 2.3 is a stationary point ($\nabla V(\boldsymbol{X_s}) = 0$) at the crossing between two adjacent basins of attraction. At this point, the PES reaches local maximum in only one direction and local minima in all other directions; thus, it is a first order saddle point. The process of locating a first order saddle point on the PES is called the transition structure search.

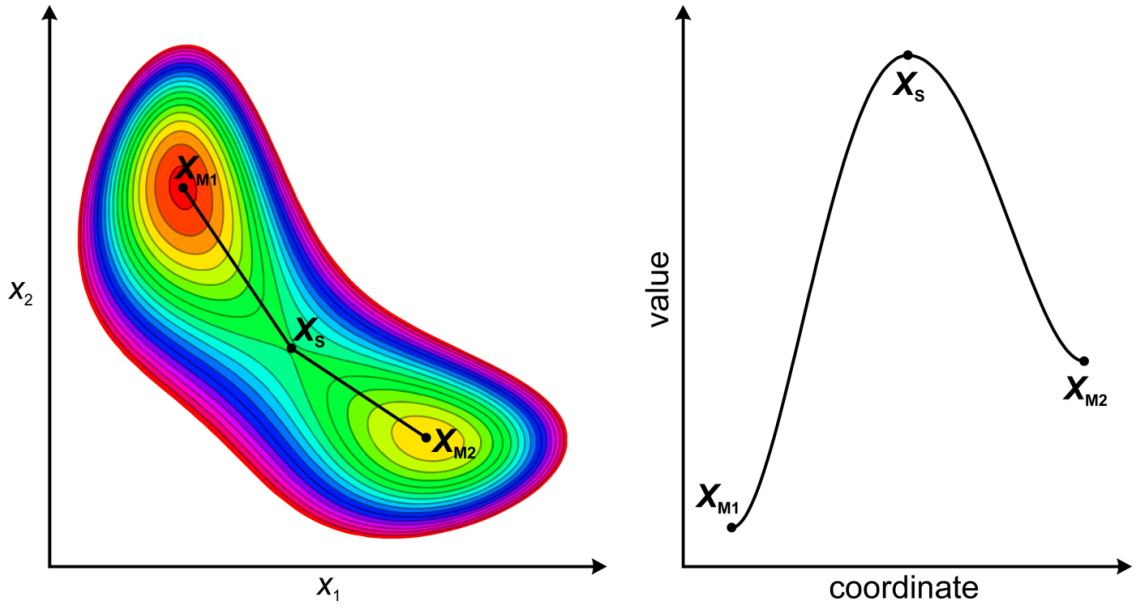

Figure 2.3: An example of PES for a reaction for which reactants and products occupy adjacent basins of attraction (left) and the energy profile associated with the selected reaction path (right). $\boldsymbol{X}_{\mathbf{M1}}$ and $\boldsymbol{X}_{\mathbf{M2}}$ denote minima, $\boldsymbol{X}_{\mathbf{S}}$ indicates the first order saddle point.

2.3.2.1 Following the reaction coordinate

The reaction coordinate is an abstract representation of the reaction progress. In terms of the underlying PES, it is a function of many variables (*e.g.* bond distances, angles) that connects the minima corresponding to the configurations of the system before (reactants) and after (products) the reaction. Following the reaction coordinate is one of the methods for finding a transition state. In this approach, a single coordinate or a combination of a few coordinates is selected as the reaction coordinate. The reaction path is constructed by a successive advance in the reaction coordinate with all remaining degrees of freedom optimized at each step (constrained optimization). The configuration with the highest energy amongst the optimized structures is an approximation to the transition structure.

2.3.2.2 Synchronous transit method

In the linear synchronous transit method (LST),[4] the continuous pathway is obtained by linear interpolation of internuclear distances:

$$\boldsymbol{D}_{\mathbf{I}}(r) = (1-r)\boldsymbol{D}_{\mathbf{R}} + \boldsymbol{r}\boldsymbol{D}_{\mathbf{P}}, \qquad (2.3.2)$$

where r is the interpolation parameter and $\boldsymbol{D}_\mathbf{I}$, $\boldsymbol{D}_\mathbf{R}$, and $\boldsymbol{D}_\mathbf{P}$ matrices contain the interpolated, reactant, and product internuclear distances, respectively. The path derived from the $\boldsymbol{D}_\mathbf{I}$ matrix connects two selected minima with a straight line (Figure 2.4). The point contained in the LST pathway characterized by the highest energy, $\boldsymbol{D}_\mathbf{I}^\mathbf{max}$, is an upper bound to the transition state energy. This guess can be further refined according to the quadratic synchronous transit (QST) methodology.[4] First, the LST maximum $\boldsymbol{D}_\mathbf{I}^\mathbf{max}$ is optimized in the direction perpendicular to the pathway. The result of such a constrained optimization process, $\boldsymbol{D}_\perp^\mathbf{min}$, is a lower bound to the transition state energy (Figure 2.4). Second, having tree non-linearly arranged points created from the $\boldsymbol{D}_\mathbf{R}$, $\boldsymbol{D}_\mathbf{P}$, and $\boldsymbol{D}_\perp^\mathbf{min}$ matrices, a curvilinear QST pathway that joins all these points can be constructed. The QST path maximum is another, presumably better, upper bound to the energy of the transition structure.

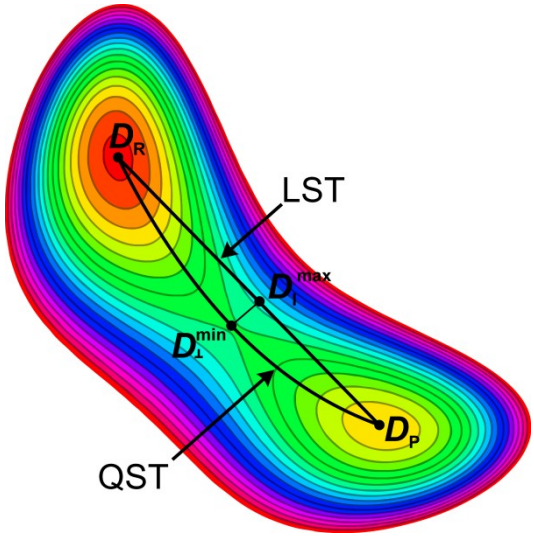

Figure 2.4: Comparison of the LST and QST pathways. $\boldsymbol{D}_\mathbf{r}$ and $\boldsymbol{D}_\mathbf{p}$ are the points corresponding to two matrices representing local minima. $\boldsymbol{D}_\mathbf{I}^\mathbf{max}$ corresponds to the matrix that represents the highest point of the LST line. $\boldsymbol{D}_\perp^\mathbf{min}$ reflects the matrix obtained after constrained optimization started from the $\boldsymbol{D}_\mathbf{I}^\mathbf{max}$ (see the text). Adapted from Ref. 4.

2.3.2.3 Nudged elastic band method

The nudged elastic band method (NEB)[6] is another method for finding saddle points and a reaction path joining two known minima. In this method, a series of structure models (images, replicas) is constructed along the path defined by two minima. Then, this whole population of images is simultaneously optimized under the constraint that each two neighbouring images are interacting with each other with a harmonic potential, i.e. are connected with a *virtual*

spring. In the course of the optimization process, the elastic band formed of these structure replicas drifts towards the minimum energy pathway. Since in this algorithm the (continuous) minimum energy pathway is approximated by a discrete set of points, the accuracy of the NEB method depends on the number of replicas used for calculations. The images characterized by the highest energies are estimates of the transition state.

2.4 Global exploration of energy landscapes

A common feature of the local exploration techniques described in the previous Section is that the exploration of the energy landscape is performed in relatively close vicinity of the given starting point(s). The aim of the global exploration techniques is to inspect multitude of basins of attraction and in this way to perform more extensive investigation of the PES topology. A brief overview of techniques used for a global exploration of the energy landscapes is presented below.

2.4.1 Educated guess approach

Educated guess in the classical *scientific method* is a technique of formulating a *hypothesis* on the basis of knowledge and experience. The applicability of this method is strongly system-dependent. Namely, if the properties of already investigated systems are transferrable to the currently studied system, an educated guess is a very reasonable starting point for further study. In the paper on investigation of the structure of an ultrathin silica film grown on the Mo(112) surface,[7] the authors declare that the constructed structure models were "inspired by the structures of layered silicates." One of these structure models was proven to be in very good agreement with the experimental data (*cf*. Chapter 3). Thus, this investigation is an example of a successfully applied educated guess. On the other hand, if the transferability of the properties is poor, the educated guess might lead to a *false hypothesis*, which in the case of atomic structure prediction means incorrect structure estimate. An example of the false structure determination based on an educated guess is the case of the $(Al_2O_3)_4$ cluster. Initially, the structure of this cluster was assumed to resemble the structure of bulk α–Al_2O_3 (an educated guess). Later, the investigation using the genetic algorithm technique yielded more stable structure which properties are in much better agreement with the experimental data. Detailed history of this problem together with corresponding sources is presented in Ref. 8.

2.4.2 Molecular dynamics method

The molecular dynamics (MD) method is based on the deterministic approach in which atomic coordinates advance along a trajectory defined by the forces acting on nuclei. In the classical molecular dynamics approach, the forces can be obtained using any chosen method that evaluates the potential energy surface on a desired level of description, *e.g. ab initio*, but the movement of nuclei is described according to the classical Netwon's laws of motion. Since the propagation is performed in discrete time steps rather than continuously, different schemes for updating the nuclei positions (propagators) exist. According to the second Newton's law of motion, the movement trajectory of the given particle is affected by the acting force which manifests itself through acceleration. Thus, for the basic Verlet propagator:

$$\boldsymbol{X}_{t+\Delta t} = 2\boldsymbol{X}_t - \boldsymbol{X}_{t-\Delta t} + \boldsymbol{a}_t(\Delta t)^{\boldsymbol{2}}. \tag{2.4.1}$$

The $t - \Delta t$, t, and $t + \Delta t$ indicies represent past, current, and future step, respectively. $\boldsymbol{X}$ is a vector of nuclear coordinates, and $\boldsymbol{a}$ represents the accelerations computed for each nucleus according to the second Newton's law of motion:

$$\boldsymbol{a}_t^{\alpha} = \frac{\boldsymbol{F}_t^{\alpha}}{M^{\alpha}}. \tag{2.4.2}$$

$\boldsymbol{F}_t^{\alpha}$ is the force acting at the time t on the nucleus α of mass M^{α}. The velocities of nuclei at the time t, $\boldsymbol{v}_t$, are calculated according to the central difference scheme:

$$\boldsymbol{v}_t = \frac{\boldsymbol{X}_{t+\Delta t} - \boldsymbol{X}_{t-\Delta t}}{2\Delta t}. \tag{2.4.3}$$

Other versions of the Verlet propagator exist as well. For example, in the leapfrog modification, the velocities and positions are calculated on time grids shifted with respect to each other by a half of the Δt period:

$$\boldsymbol{X}_{t+\Delta t} = \boldsymbol{X}_t + \boldsymbol{v}_{t+\frac{\Delta t}{2}}\Delta t, \tag{2.4.4a}$$

$$\boldsymbol{v}_{t+\frac{\Delta t}{2}} = \boldsymbol{v}_{t-\frac{\Delta t}{2}} + \boldsymbol{a}_t\Delta t. \tag{2.4.4b}$$

Since the movement of nuclei is described according to the classical Netwon's laws of motion, the sum of potential and kinetic energy is conserved throughout the simulation. Thus, the potential energy surface is sampled within the microcanonical ensemble. The canonical ensemble sampling can be imposed by introducing a thermostat, *i.e.* an algorithm that regulates the kinetic energy of the system. Theoretical aspects of the thermostat algorithms are covered in Ref. 9.

2.4.3 Metropolis Monte Carlo technique

Contrary to the deterministic nature of molecular dynamics, the Metropolis Monte Carlo technique belongs to the group of stochastic methods. In the Metropolis Monte Carlo approach, a change of the state of the system occurs according to an event selected from a set of predefined events, such as translation of a certain atom or rotation of the whole group of atoms. If $\hat{A}_i$ is an operator that represents an event i, it can be selected from a set of predefined events $G = \{\hat{A}_i\}$ with the probability $p(\hat{A}_i)$. Such an event is then applied on the system state $\boldsymbol{X}_n$, transitioning it to the state $\boldsymbol{X}_n^i$:

$$\hat{A}_i(\boldsymbol{X}_n) = \boldsymbol{X}_n^i. \tag{2.4.5}$$

This transition is accepted with the probability $p(\boldsymbol{X}_n \to \boldsymbol{X}_n^i)$. If the transition (2.4.5) is accepted, the next iteration of the algorithm starts from the new state $\boldsymbol{X}_{n+1} = \hat{A}_i(\boldsymbol{X}_n) = \boldsymbol{X}_n^i$. Otherwise, the next iteration starts from the old state $\boldsymbol{X}_{n+1} = \boldsymbol{X}_n$.

Since in the thermal equilibrium the distribution of states over the energy is described by the Maxwell–Boltzmann statistics, for the molecular modelling applications the probability $p(\boldsymbol{X}_n \to \boldsymbol{X}_n^i)$ is chosen to mimic this distribution:

$$p(\boldsymbol{X}_n \to \boldsymbol{X}_n^i) = \begin{cases} 1, & \Delta E \leq 0 \\ e^{-\frac{\Delta E}{kT}}, & \Delta E > 0, \end{cases} \tag{2.4.6}$$

with k being Boltzmann constant, T temperature, and

$$\Delta E = E_{\boldsymbol{X}_n^i} - E_{\boldsymbol{X}_n} \tag{2.4.7}$$

the difference in energy between tested and initial state. According to Eq. 2.4.6, for $T = 0$ only downhill and energetically neutral steps are accepted, while the higher the temperature, the easier it becomes to accept the steps that increase the energy of the system.

2.4.4 Genetic algorithms

A genetic algorithm is a search heuristic that employs principal mechanisms of natural evolution to simulate a time development of a whole set of trial solutions. Since over the years the genetic algorithms have been applied in many fields of research, *e.g.* chemistry,[10] psychology,[11] oncology,[12] or engineering,[13] various versions of the algorithm, tailored to particular needs, exist.

2.4.4.1 Genesis of genetic algorithms

In 1957, Fraser published the first paper of the series "Simulation of Genetic Systems by Automatic Digital Computers."[14] In this series, the author demonstrates that a digital computer is a suitable tool for simulating the genetic operations (*e.g.* reproduction). Fraser indicates that (i) the digital computer inherently operates on binary representation of data and that (ii) the genetic information can be easily stored in a binary form. Thus, by taking advantage of the native support for logical and algebraic operations, "the sequences discussed above, when formulated in a programme suitable for an automatic computer, allow simulation of genetic systems."[14] Subsequently, the artificial evolution techniques were used to solve certain problems. A review of these early approaches is given in Chapter 4 of the textbook "Genetic Algorithms in Search, Optimization and Machine Learning" by David E. Goldberg.[15] Out of the examples presented there, two works are of particular importance. The first one is the monograph "Adaptation in Natural and Artificial Systems"[16] by J. H. Holland, and the second one is K. A. De Jong's dissertation "An Analysis of the Behavior of a Class of Genetic Adaptive Systems."[17] Both will be shortly discussed here.

In the monograph "Adaptation in Natural and Artificial Systems"[16] (1975), Holland investigates the role, the mechanisms, and the properties of the *adaptation* process. The process of evolution of natural organisms is shown as an example of an optimization procedure. The natural evolution is performed in a vast space of possible combinations of alleles, in which genes cannot be separately optimized due to unpredictable coupling effects (*i.e.* the nonlinearity of the performance with respect to a change in the genetic code). The

author points out that the given biological organism can be very successful in a certain environment while under different conditions it can fail. On the other hand, many organisms can fit to the given environment, although their genetic code differs significantly. Thus, on Page 11 the author asks: "[…] the central question for genetic systems is: How are initially unsuited structures transformed to […] structures suited to a variety of environmental niches?" Starting from these biology-driven considerations, Holland proposes a general scheme which can be applied to a variety of optimization problems. The scheme was originally called *genetic plan* but the name later evolved to *genetic algorithm*.

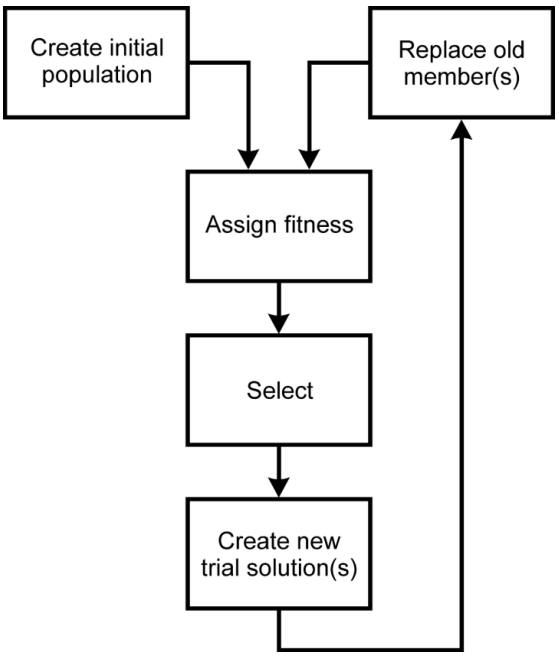

Figure 2.5: The genetic plan proposed by Holland.

The genetic plan introduced in Holland's work (Figure 2.5) is a search algorithm, which operates on a *population* (set) of *structures* (trial solutions to the given problem). The search procedure is an iterative process that starts with an *initial population* (initial set of trial solutions). Each of the trial solutions is characterized by *observed performance* (fitness). The performance of an individual determines its probability to create offspring – the better the performance, the higher the probability. The offspring is created by application of one of the *genetic operators* on a copy of the parent. In the minimal scheme these are: *inversion*, *mutation*, or *crossover*. The inversion operator modifies the sequence in which data is represented in the genetic code, whereas the mutation operator randomly changes an arbitrary part of the genetic code. While these two operators act on a single structure, the crossover operator is a binary operator that requires selection of a mating partner. During the crossover,

representations of two parent structures are recombined – genetic codes of both parents are divided into two parts and the offspring's genetic codes are created by joining the complementary parts. Independently on the method in which the offspring was created, a single new trial solution replaces one element of the population. Holland explicitly mentions two classes of algorithms which differ by the number of structures replaced in the population in a single cycle. In one case, let us call it H1, only one randomly selected member of the population is replaced per cycle. For the other case, H2, the cycle involves exchange of the whole population.

Holland proposes the genetic scheme having in mind very complex optimization problems. Already *a simple artificial adaptive system*, given in the book as an example, is a pattern recognizer with 100 pixels and 10^{100} distinct ways to distribute ten possible weight values among the pixels. The next considered example is *a complex natural adaptive system* with 10^{3000} possible combinations. Although an enumerative (exhaustive) search guarantees finding the optimal solution to these problems, time and resources are two obstacles that prohibit using such an approach. As the author writes on Page 17: "If 10^{12} structures could be tried every second […], it would take a year to test about $3 \cdot 10^{19}$ structures, or a time vastly exceeding the estimated age of the universe to test 10^{100} structures". Holland's scheme does not guarantee finding the optimal solution to the given problem since it is not designed to ensure exploration of the whole solution space. The key feature of the genetic plan is its *robustness*, intuitively understood as the ability to find *acceptable* solutions in *acceptable* time.

In 1975, K. A. De Jong published his dissertation "An Analysis of the Behavior of a Class of Genetic Adaptive Systems."[17] In this work, De Jong performs a deep analysis of the influence of the algorithm parameters on the efficiency of the search. The test environment for his study was defined by five functions: (i) F1 – a simple three-dimensional parabola; (ii) F2 – the Rosenbrock function;[18] (iii) F3 – a five-dimensional step function; (iv) F4 – a 30-dimensional quartic function with Gaussian noise; and (v) F5 – a two-dimensional multimodal function:

$$\frac{1}{f(x_1, x_2)} = \frac{1}{500} + \sum_{j=1}^{25} \frac{1}{j + \sum_{i=1}^{2} (x_i - a_{ij})^6}, \tag{2.4.8}$$

with constants a_{ij} given in the dissertation.

De Jong introduced the aspect of a generation gap, *i.e*. coexistence of the elements that belong to different generations. The extent of the overlap is determined by the *generation gap parameter* G, $0 < G \leq 1$. If N is the total size of the population, then $G * N$ is the number of created child structures that replace the same number of structures in the population and $(1 - G) * N$ is the number of the elements of the old population that survive. The Holland's schemes described earlier are two special cases of generation overlapping, the H1 scheme with $G = \frac{1}{N}$ (one element replaced per generation) and the H2 model with $G = 1$ (all elements replaced per generation). Nowadays, the schemes that use relatively small *G* are called the steady-state genetic algorithms.[19]

De Jong considered six different variants of the genetic plan: (i) R1 – Holland's scheme H2; (ii) R2 – *elitist model* – similar to the R1 model but the best trial solution is always preserved in the population; (iii) R3 – *expected value model* – similar to the R1 scheme, however the fitness is translated directly into number of children created by a particular structure (not into probability); (iv) R4 – *elitist expected value model* – combination of the R2 and R3 models; (v) R5 – *crowding factor model* – a member of population to be replaced by the given child is selected considering similarity to the child; and (vi) R6 – *generalized crossover model* – a more general rule for choosing crossover points.

The influence of population size, mutation probability, and generation gap parameters on the efficiency of the algorithm was investigated for the most basic plan, namely R1, applied to the simplest problem F1. The parameters were analyzed with respect to a long-run convergence and the speed of convergence. The result of this investigation can be summarized as follows: (i) larger populations maintain more gene diversity and thus are superior in the long-run convergence; (ii) smaller populations are characterized by lower inertia and thus converge faster (although not necessarily to the best solution); (iii) the larger the mutation rate the better the gene diversity; however, (iv) too high mutation rate hinders the long-run convergence (too high randomness in the search procedure); (v) the larger population overlap the better initial convergence of the algorithm; however (vi) the performance gain is accompanied by loss of gene diversity, which in turn leads to premature convergence of the algorithm.

In one of the variants, the R5 scheme, De Jong proposes a way to maintain genetic diversity in the population without overusing the mutation operator. The scheme requires the populations to overlap, *i.e*. the generation gap parameter $G < 1$. In addition, a new parameter – the crowding factor CF – is introduced. CF determines the number of population members

randomly preselected as candidates for the replacement with the new trial candidate. Then, the similarity assessment between each element of this subset and the new candidate solution is performed. The child replaces the most similar individual. It is a generalization of the R1 scheme since for $CF = 1$ only one candidate is randomly selected and no similarity assessment is required. On the other hand, for $CF = N$, the replacement is based only on the similarity; hence, the result is purely deterministic.

All six genetic plan variants R1 – R6 were tested for all five environments F1 – F5. De Jong's investigations performed for only limited number of test functions already show that: (i) the efficiency of the genetic algorithm depends on many factors; (ii) the optimal set of parameters is case-specific; and (iii) locating the global minimum is not guaranteed. These drawbacks clearly discourage selecting the genetic algorithm as a method for finding the global minimum for simple functions. On the other hand, the same tests show that the genetic scheme is much more robust than the random search. This observation is of extreme importance in the light of Holland's considerations of complex optimization problems which reach 10^{100} possible trial solutions.

Goldberg's textbook,[15] published almost 15 years after the works of Holland and De Jong, lists more than 80 publications in which the genetic algorithm was applied to various optimization problems. Goldberg describes also many ways in which the genetic algorithm can be modified. Among these are: (i) using non-binary coding; (ii) creating a hybrid scheme in which the genetic algorithm scans the solution space and locally convergent scheme is used for local optimization; (iii) introducing a penalty function.

2.4.4.2 Genetic algorithms applied to structure determination problem

In the previous Section, I described the foundations of the genetic algorithms. During the genetic algorithm run, the population of trial structures adapts to the given environment. Similar mechanisms of natural evolution are applied to the genetic scheme itself. The genetic algorithm (a single element of population) is tailored to certain applications (environment). The ideas used in one algorithm instance are crossed-over with the ideas used by other groups. In this way, the initial genetic scheme described by Holland[16] forks to become a whole family of algorithms connected by the idea of artificial evolution. In Goldberg's textbook[15] published in 1989, more than 80 applications of the genetic algorithms are listed. A

web of science[20] search performed 25 years later returns already 1300 papers published *only* in the year 2013 that contain the "genetic algorithm" phrase *in the title*, not to mention those containing this phrase as a keyword. This multitude of publications makes it virtually impossible to review all of them. Thus, in this Section, Ifocus on evolution of the genetic algorithms applied only to the problem of determination of atomic structures.

The early applications of the genetic algorithms to structure determination used rather straightforward implementations of Holland's scheme. Blommers, Lucasius, Kateman, and Kaptein[21] applied the genetic algorithm to determine the structure of a pyrimidine photodimer. The exploration was restricted to a part of the conformational space, *i.e.* only eleven torsion angles were varied. The conformations were encoded in a binary form with angles represented as integers. The authors used the Gray integer representation[22] as more natural for the problem. This decision was justified in a following manner: "Gray integers have the property that a change in any of its bits yields a decoded real value that is close to the original one" (p. 46) and further "for torsion angles the values 0 and 360 are close to each other in Gray space" (p. 46). The fitness function was a measure of agreement between the experimentally determined interatomic distances and the distances obtained for the generated structure model. The authors note that, in the absence of additional control mechanisms, the population could be easily dominated by a set of similar conformations. The fittest structure models obtained from the GA runs were locally optimized.

Judson, Colvin, Meza, Huffer, and Gutierrez[23] applied the genetic algorithm to a "2-D polymer", which is an idealized chain formed of "atoms" connected with the nearest neighbours by rigid "bonds". The energy of the polymer was calculated from the Lennard-Jones pairwise interaction for all atoms. The relative arrangement of atoms (conformation) is determined by the set of angles formed by adjacent bonds. Similarly to the implementation discussed in the previous paragraph, the genes contained information about the angles encoded with the Gray integer representation.[22] The authors used De Jong's elitist model (R2),[17] *i.e.* the best species was always preserved in the population. In addition, the authors applied the *Lamarckian operator*. This is a unary operator, which generates a locally optimized copy of the given species and replaces the original individual with its optimized version. The offspring generated from such an optimized individual obtains "traits learned by the parent" (p. 282). This property is in accordance with the idea of inheritance of acquired characteristics, *i.e.* Lamarckism, hence the name of the operator.

Smith[24] applied the genetic algorithm to a problem of binary alloy models. In his implementation, the probability of becoming a parent was proportional to the weight obtained using a dynamically scaled exponential scheme:

$$w_i = e^{-\alpha \frac{E_i - E_{\min}}{E_{\text{aver}} - E_{\min}}}, \tag{2.4.9}$$

where w_i is the weight of the *i*th species, E_i the energy of the *i*th species, $E_{\min}$ the lowest energy in the population, E_{aver} the average energy of the population, and α the scaling factor. The author notes that the scaling factor α "controls the relative weight of the best solution as compared to that of the average" (p. 136). Indeed, for $\alpha = 0.0$, the weights are distributed uniformly, whereas higher α values bias the relative weight towards the low energy species (Figure 2.6). In addition, Smith used the generation gap concept[17] which allows partial replacement of the population per generation (the generation gap / generation overlap idea is explained in more detail in Section 2.4.4.1). In order to perform this partial replacement, Smith eliminates "the poorest"[24] elements of the old population, *i.e.* applies an energetic criterion.

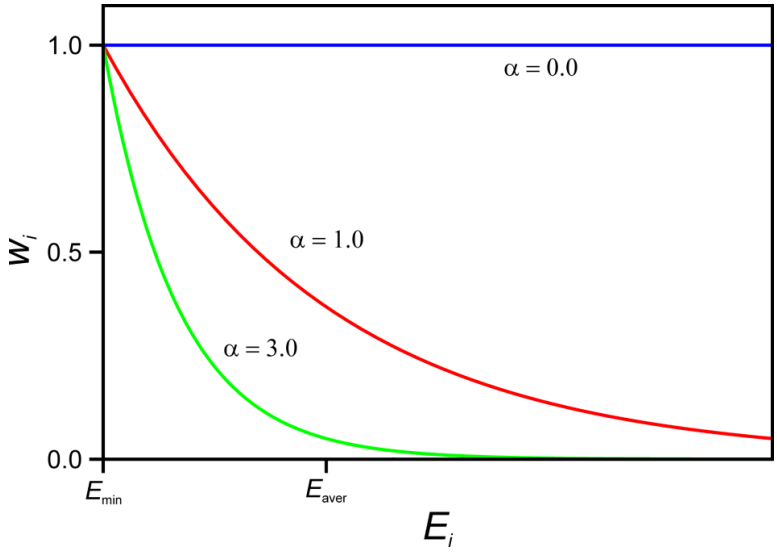

Figure 2.6: Visualisation of the $\boldsymbol{w_i(E_i)}$ weighting function, defined in Eq. 2.4.9. The selected values of the scaling factor $\boldsymbol{\alpha}$ are indicated.

Hartke[25] used the genetic algorithm to perform a global optimization of the Si_4 cluster structure model. The cluster was optimized on an empirical potential energy surface; however, the author states that the final goal is to use an *ab initio* PES. Hartke used an exponential

fitness assessment which was fixed to return 1.0 weight for the known global minimum and 0.01 for structures characterized by the given relatively high energy value. This is different from the dynamically scaled exponential scheme discussed in the previous paragraph. The configurations were defined using six degrees of freedom. The values were encoded in binary strings but, contrary to the studies described earlier, these strings represented real numbers, not integers.

McGarrah and Judson[26] analyzed performance of the genetic algorithm on an example of the cyclic hexaglycine molecule. In their paper, the authors discuss the efficiency of hybrid global-local minimization schemes, *i.e.* the schemes composed of global and local optimization procedures. The authors point out that the local optimization is performed much more efficiently with the specialized local optimization schemes than with the genetic algorithm. Thus, "the GA has the objective of generating as many different good candidate initial solutions for the gradient minimization as possible and of spending the minimum amount of time searching within any single well."[26] The authors also note that convergence of the genetic algorithm population to a single basin of attraction is undesirable since the local optimization of such species would result in a set of virtually identical structure models. On the other hand, an ideal genetic algorithm run should produce a population containing conformations which represent different local minima. In fact, the members of population that lead to the same local minimum are called by the authors "redundant" since they "provide no new information regarding minimized conformations".[26]

Brodmeier and Pretsch[27] tested genetic algorithm performance for conformational analysis of the *n*-decane and 3-methylnonane molecules. Only the torsion angles were varied but the full structure was used for the energy (fitness) evaluation. For assessment of fitness, the authors used the dynamically scaled exponential scheme presented in Eq. 2.4.9. The generation gap concept was used (*cf.* Section 2.4.4.1). The authors used the genetic algorithm as a tool to explore the PES in order to find reasonable structure models that can be later optimized (prescreening method). Similarly to McGarrah and Judson,[26] Brodmeier and Pretsch aimed at obtaining the population formed of species belonging to different basins of attraction. In order to prevent premature convergence of the genetic algorithm run, the authors used a similarity recognition tool. Each new trial solution was compared with the elements of the population. The similarity analysis was based on the differences of torsion angles. If the sum of the absolute differences was smaller than the given threshold, an additional mutation of the trial solution was enforced.

Deaven and Ho[28] applied the genetic algorithm to carbon clusters of size up to C_{60}. The authors did not use the conventional binary string coding. Instead, the configurations were directly represented using the Cartesian coordinates. The crossover operator was modified to act on such a 3D representation in a following manner. First, a random plane was selected and aligned to pass by the centres of mass of both parent structure models. Second, both parent clusters were cut by this plane. Then, two children were created by joining the complementary parts of both clusters. f the atomic composition of the children was not the same as that of the parents, both parent structure models were translated in the opposite direction from the cut plane until the correct composition was obtained. The authors implemented two real space mutation operators. One of them was a random translation of a single atom. The second one performed a gradient-guided search in order to locate a neighbouring basin of attraction. Since mating and mutation procedures are not performed on the coded representation of the structure (*i.e.* genetic code, genotype) but rather on the structure directly (*i.e.* manifestation of the traits, phenotype), this approach was later referred to as the *phenotype evolution* or the *phenotype algorithm*.[29, 30] Deaven and Ho applied local optimization to each obtained trial structure. Such an optimized species was then tested against the members of population. The child replaced a member of population only if it was energetically more stable than that member. The authors prevented the premature convergence of the genetic algorithm run using the energy criterion, *i.e.* the structural diversity was maintained by enforcing the requirement of a minimum energy difference on all members of population.

Curotto, Matro, Freeman, and Doll[31] performed global optimization of the Ni_n, n = 2-13 clusters. Their implementation of the genetic algorithm was similar to that of Deaven and Ho.[28] The most significant difference (from the perspective of this dissertation) lies in the structural diversity algorithm. In the work of Deaven and Ho, the structure was not allowed to enter the population if its energy was too close to the energy of any other member of the population. Curotto *et al*. noted that certain isomers are characterized by very similar energies and thus very tight (*i.e.* costly) local optimization would be required to distinguish between two different structures using only the energy criterion. As a solution to this problem, the authors applied a topology-based scheme, *i.e.* a similarity recognition tool. The similarity was determined "by attempting to superimpose each child-child and child-parent pair with a set of translation-rotation operations".[31] For each two structures selected for comparison, one of the structures was used as a template. The first atom from this structure was selected as an "origin", which together with the second atom defined an axis, and these two atoms together

with the third one defined a plane. Similarly, origins, axes, and planes were created for the second structure with a difference that not one but all possible combinations were obtained. Thus, if *N* denotes number of atoms in the second cluster, there were *N* origins to select, for each of those (*N*-1) axes to create, and for each of these *N*(*N*-1) combinations (*N*-2) possible planes, which in total gives *N*(*N*-1)(*N*-2) origin-axis-plane combinations. For each of those combinations, the second structure was translated and rotated in the way that the origins, axes, and planes of two clusters were aligned and the positions of the atoms in both structures were compared. If the differences were smaller than the given threshold, the structures were considered similar. In this way, the authors were able to distinguish degenerate structures, *e.g*. optical isomers.

Hartke[30] applied a phenotype algorithm similar to that of Deaven and Ho[28] to the set of the Lennard-Jones clusters of the size up to 150 atoms. In his implementation, the advance to the next generation was performed using an intermediary extended population. Such an extended population was formed from all members of the current population together with the optimized children structures. Then, the elements of the extended population were sorted according to their energy and the least stable species were removed to restore the original population size. In this way, at any moment of the algorithm run the population contained only the best solutions discovered (the strong elitist strategy). Structural diversity was maintained using two algorithms. The first one was the energy-based scheme used by Deaven and Ho[28] and the second one was based on structural similarity.

Finally, the genetic algorithms were also used to obtain structural models of periodic systems. Bazterra, Ferraro, and Facelli[32] performed global optimization of crystal structures of the benzene, naphthalene, and anthracene molecules. The authors optimized the crystallographic axes and spatial arrangement of the molecules in the unit cell, *i.e*. molecular positions (centres of mass), and Euler angles. On the other hand, the molecules themselves were treated as rigid bodies. Chuang, Ciobanu, Shenoy, Wang, and Ho[33] applied the genetic algorithm to the surface structure determination problem. Although Chuang *et al*. investigated a 2D periodic system, their implementation had many features in common with the algorithm of Deaven and Ho[28] used for cluster optimization. The shared ideas included: i) Cartesian representation of atomic positions; ii) local optimization of the trial structures; iii) replacement of the elements of the population only if the tested child is characterized by lower energy; iv) real-space crossover and mutation operators (phenotype algorithm). Chuang *et al*. maintained the structural diversity of the population using a two-step similarity recognition scheme. In the

first step, the surface energies of the structures were compared. If the difference of the surface energies was smaller than the given threshold, the structures were flagged as possibly similar. In the second step, the structures characterized by similar surface energies were analyzed for topological similarity. Two structures were considered equal when the interatomic distances (up to the given threshold) matched for both of them. For the purpose of the real-space crossover operator, both parent structures were divided into two parts Figure 2.7. The "active" part was defined as the topmost 500 pm layer of atoms. The rest, *i.e*. the bottom part, was treated as bulk-like (inert) and thus ignored. The two active parts obtained from both parents were then cut by the same random plane and the complementary sections were assembled into the new active part. This new recombined layer was then joined with the bulk-like template support and subsequently optimized.

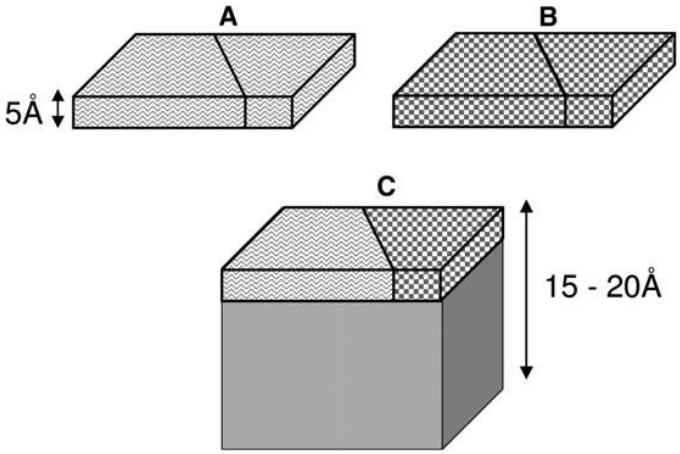

Figure 2.7: The mating operation used by Chuang *et al*. Reprint from Ref. 33.

2.4.5 Simulated annealing scheme

Annealing is a technique of heat treatment during which the material is heated above its critical temperature and slowly cooled afterwards. This process is used in metallurgy and materials science to overcome thermodynamic barriers that prevent material from acquiring the desired structure. The annealing process became inspiration to modify the Metropolis Monte Carlo algorithm so that it *simulates* the *annealing* mechanism by first "melting" and then "freezing" the system.[34] Since in the Metropolis Monte Carlo technique the temperature is a parameter that governs the probability of accepting uphill steps (Eq. 2.4.6), the initial increase of the temperature allows overcoming even large energy barriers. In the second phase, the cooling is performed. This traps the system in a certain area of the PES. Cooling schedule is crucial for the efficiency of the simulated annealing procedure. Quenching freezes

the system in the current basin of attraction, whereas a slow decrease of temperature progressively diminishes the probability of climbing high barriers. The second scheme permits further exploration of the PES and eventually gives higher probability of locating a deeper local minimum. The simulated annealing technique can be applied as an extension also to other temperature-based algorithms, such as molecular dynamics (with a thermostat).

2.4.6 Basin hopping scheme

Basin hopping is a search technique based on the transformation of the energy landscape.[35-37] In this method, the energy of any structure is mapped onto the energy of the associated local minimum, *i.e.* onto the lowest energy obtainable in the given basin of attraction. If $\boldsymbol{X_n}$ represents a particular configuration and $E(\boldsymbol{X_n})$ a value of the potential energy for that configuration, then the transformed energy for that configuration, $\tilde{E}(\boldsymbol{X_n})$, is given by:

$$\tilde{E}(\boldsymbol{X_n}) = \min\{E(\boldsymbol{X_n})\} = E(\boldsymbol{X_{\mathbf{min}}}), \tag{2.4.10}$$

where $\min\{E(\boldsymbol{X_n})\}$ denotes the energy returned by a local optimization procedure starting from the configuration $\boldsymbol{X_n}$ and $\boldsymbol{X_{\mathbf{min}}}$ stands for associated local minimum configuration. Since

$$\tilde{E}(\boldsymbol{X_n}) = E(\boldsymbol{X_{\mathbf{min}}}) \leq E(\boldsymbol{X_n}), \tag{2.4.11}$$

the energies for all local minima are invariant under this transformation. The transformation (2.4.10) is formally performed for all configurations, thus the whole transformed potential energy surface contains only plateaus and discontinuities (Figure 2.8). Note that virtually identical transformation of the PES is performed by application of the Lamarckian operator within the genetic algorithms (Section 2.4.4.2).

The term *basin hopping* originates from the idea that exploration of an energy landscape is performed by *hopping* directly between different *basins*, which are represented by corresponding minima located in the course of local optimization of trial configurations. The basin hopping method requires a local optimization algorithm to perform the step presented in Eq. 2.4.10 but also a global exploration technique to move across the original energy landscape. Successful applications include the canonical Monte Carlo[35] and genetic algorithm[10] techniques.

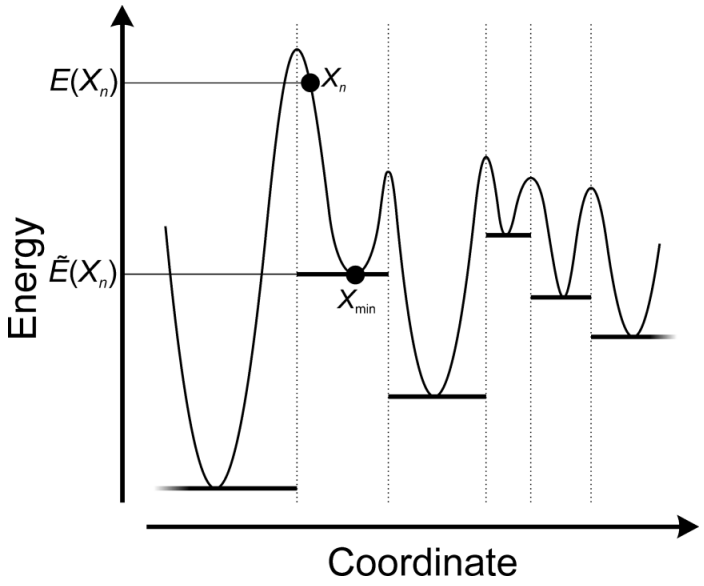

Figure 2.8: An example of the PES transformation applied in the basin hopping method.

2.4.7 Particle swarm optimization technique

This Section is based on the review article by R. Poli.[38] The particle swarm optimization technique is an iterative process that operates on a set (*swarm*) of trial solutions (*particles*) to the given problem. Each particle advances within an individual trajectory whilst in each step evaluating the loss function (*e.g.* energy) at the current position. The particles are grouped in subsets called *social neighbourhoods*. Each particle has an independent memory of the visited points with particular significance of the point characterized by the lowest value of the loss function. Such a point for the ith particle will be henceforth denoted as $\boldsymbol{P}^i$. Analogously, the point with the lowest value of the loss function for the whole social neighbourhood will be denoted as $\boldsymbol{S}^x$ where x is any particle that belongs to the given social neighbourhood. If $\boldsymbol{v}_{n-1}^i$ denotes the velocity in the search space obtained in the previous iteration for the ith particle positioned at $\boldsymbol{X}_{n-1}^i$, then the new velocity, $\boldsymbol{v}_n^i$, is given by:

$$\boldsymbol{v}_n^i = \alpha \boldsymbol{v}_{n-1}^i + \beta R1_n^i \left(\boldsymbol{P}^i - \boldsymbol{X}_{n-1}^i\right) + \gamma R2_n^i \left(\boldsymbol{S}^i - \boldsymbol{X}_{n-1}^i\right), \tag{2.4.12}$$

where α is the *inertia weight*, β and γ are *acceleration coefficients*, and $R1_n^i,\ R2_n^i \in [0,1]$ are two numbers generated randomly for each particle and iteration. The acceleration terms $\beta R1_n^i \left(\boldsymbol{P}^i - \boldsymbol{X}_{n-1}^i\right)$ and $\gamma R2_n^i \left(\boldsymbol{S}^i - \boldsymbol{X}_{n-1}^i\right)$ are responsible for attracting the particle in the direction of the individual and social neighbourhood minima, $\boldsymbol{P}^i$ and $\boldsymbol{S}^i$, respectively. The

relative weight of the γ coefficient determines the strength of correlation of movements for the whole social neighbourhood. $\gamma = 0$ defines the swarm of non-interacting particles.

2.4.8 Metadynamics method

The metadynamics is a technique for exploration of the energy landscapes introduced by Laio and Parrinello in the paper titled "Escaping free-energy minima."[39] The motivation for designing this algorithm was the inefficiency of the Metropolis Monte Carlo and molecular dynamics exploration techniques in obtaining proper statistical data when these methods are applied to problems that involve exploration of the potential energy surface in regions that incorporate relatively large barriers, *e.g*. related to a chemical reaction.

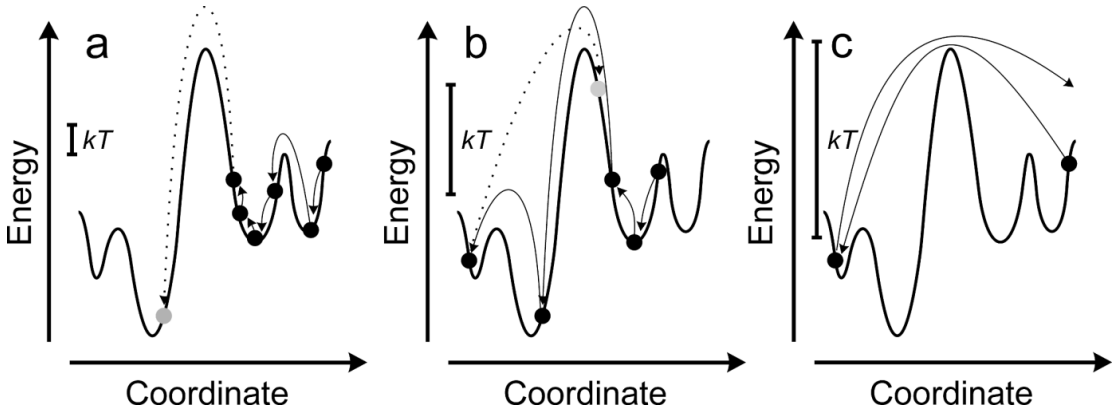

Figure 2.9: Illustration for the idealized molecular dynamics simulations performed with different temperature *T* on the energy landscape that contains a single relatively high energy barrier, which separates regions of relatively small energy barriers. The black points connected with solid arrows represent relatively probable steps whereas the dashed arrows and grey points represent the steps that have very low occurrence probability.

The difficulty connected with exploration of such energy landscapes is illustratively explained in Figure 2.9. Figure 2.9a represents the situation in which the simulation is performed at a relatively low temperature. In that case, the region neighbouring the starting point of the simulation will be explored very extensively since it contains only small energy barriers. On the other hand, the probability of overcoming the large barrier at a relatively small temperature, and thus exploration of the landscape on other side of that barrier, is extremely low. Therefore, very long simulation time would be required to at least once observe such an event. A situation in which temperature is large enough to overcome the barrier from one side is shown in Figure 2.9b. In that situation, one part of the PES will be explored only until the barrier is overcome, but then the system becomes trapped on the other side of the barrier and cannot move back. This can lead to severe bias in the final statistics, *i.e*. the statistical weight

of the region visited in the beginning of the simulation might be underestimated or overestimated, depending on the total simulation time. In the third example (Figure 2.9c), temperature is large enough to let the system spontaneously move between both sides of the barrier. In this case, unfortunately, the sampling of the energy landscape will not reproduce the fine details for the regions separated by the low energy barriers. Similar problems are encountered for the Metropolis Monte Carlo simulations.

The metadynamics method was originally developed as a modification of the molecular dynamics technique. In the metadynamics algorithm, the dimensionality of the search space is reduced with respect to the original PES to the set of so-called *collective coordinates* $\boldsymbol{s} = \{s_i\}$, $i = 1, \dots, n$. A collective coordinate s_i is a variable that is chosen as a descriptor of the system state. For example, in the original metadynamics publication, in which the authors investigated dissociation of a NaCl molecule in water, the three selected collective coordinates were: Na-Cl bond distance, the value of the electric field on the Na atom, and the value of the electric field on the Cl atom. The exploration of the reduced search space is performed according to the forces F_i^t expressed in the dimensions of the collective coordinates:

$$F_i^t = -\frac{\partial E(\boldsymbol{s})}{\partial s_i^t}. \tag{2.4.13}$$

The index t denotes tth iteration, and the $E(\boldsymbol{s})$ represents the free energy of the system, which is expressed as a function of the collective coordinates. In the original publication, the forces F_i^t were computed using an ensemble of independent replicas of the system. All images were characterized by the same value of the collective coordinates $\boldsymbol{s}^t$. Each replica was subject to a constrained molecular dynamics run, with the constraints introduced in the form of Lagrange multipliers λ_i:

$$\sum_{i=1,\dots,n} \lambda_i (s_i - s_i^t). \tag{2.4.14}$$

After all constrained runs finished, the force was obtained by averaging the respective Lagrange multipliers:

$$F_i^t = \langle \lambda_i \rangle. \tag{2.4.15}$$

The key feature of the metadynamics technique, in addition to the collective coordinates, is the history-dependent potential $V(\boldsymbol{s}, t)$:

$$V(\boldsymbol{s}, t) = W \sum_{t' \leq t} \prod_i e^{-\frac{\left|s_i - s_i^{t'}\right|^2}{2(\Delta s_i)^2}}. \tag{2.4.16}$$

The potential $V(\boldsymbol{s}, t)$ is a sum of multidimensional Gaussian functions centered at points visited in the course of the simulation, $\boldsymbol{s}^{t'}$. Δs_i governs the width of the Gaussian for the *i*th collective variable and W determines the height of the Gaussian. The $V(\boldsymbol{s}, t)$ potential is defined as repulsive. Thus, including it in Eq. 2.4.15 biases the search trajectory in the direction of not explored parts of the energy landscape:

$$F_i^t = \langle \lambda_i \rangle - \frac{\partial}{\partial s_i} V(\boldsymbol{s}, t). \tag{2.4.17}$$

Such an approach is informally described as using *computational sand* (*i.e.* the Gaussian functions) to fill the free energy wells.[40] In the limit of very long simulation time and using infinitesimally small Gaussian functions, the energy landscape can be estimated from the obtained potential:[40]

$$\lim_{t \to \infty} V(\boldsymbol{s}, t) \approx -E(\boldsymbol{s}) + const. \tag{2.4.18}$$

2.4.9 Umbrella sampling scheme

The umbrella sampling is a technique for systematic sampling of areas of the potential energy surface introduced by Torrie and Valleau.[41] Initially, the method was developed to efficiently estimate the free energy differences in a Monte Carlo experiment. In the umbrella sampling technique, the original Boltzmann distribution (*cf.* Eq. 2.4.6) is modified by a weighting function so that the sampling is shifted to the area of interest. This can be equally represented as modifying the original Hamiltion operator with restraining potentials:[42]

$$\hat{H}_{\{\lambda\}}(\boldsymbol{X}) = \hat{H}_0(\boldsymbol{X}) + \sum_{i=1}^{n} \lambda_i \hat{V}_i(\boldsymbol{X}), \tag{2.4.19}$$

with $\hat{H}_0(\boldsymbol{X})$ representing the original Hamiltonian for the system configuration $\boldsymbol{X}$ and $\hat{V}_i(\boldsymbol{X})$ is the ith restraining potential applied with a coupling parameter λ_i. Under certain conditions, the properties of the unbiased ensemble can be a posteriori retrieved from the umbrella sampling simulations.[41]

3. Educated guess used for atomic structure determination*

3.1 Introduction

In this Chapter, I present selected studies of various ultrathin silica phases performed in collaboration with experimental groups affiliated with the Chemical Physics Department of the Fritz Haber Institute of the Max Planck Society (Director: Prof. Dr. Hans-Joachim Freund). All of the atomic structure models presented in this Chapter are results of a successful application of the educated guess approach.

The atomic structure of the crystalline ultrathin silica film synthesized on the Mo(112) substrate was first identified in 2005 by Weissenrieder *et al*.[7] The structure (Figure 3.1) was determined in the course of a synergic theoretical and experimental investigation in which a multitude of structure models were computationally tested against the experimentally obtained data. The analyzed atomic structure models were "inspired by the structures of layered silicates,"[7] thus obtained using the educated guess approach.

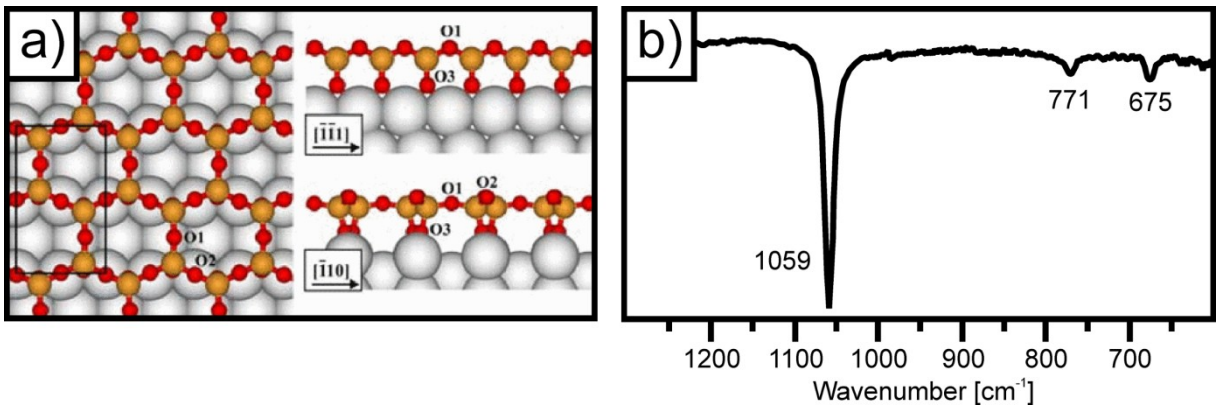

Figure 3.1: a) The most stable structure of the single-layer SiO_2/Mo(112) film proposed in Ref. 7; the Si_4O_{10} surface unit cell is indicated; Si, yellow; O, red; Mo, gray; b) the experimental IRA spectrum of the silica film grown on Mo(112). Reprints from Ref. 7.

Such an ultrathin layer grown on the metal surface can be characterized with almost atomic resolution using the real space structure imaging techniques such as the scanning tunnelling microscopy (STM) or the atomic force microscopy (AFM). On the other hand, such a thin crystalline structure deposited on a pristine metal surface (single crystal) is a convenient

* Based on Refs. W2, W3, W4, W7

model for computational studies. Thus, this system provides a unique opportunity to perform joint experimental and theoretical studies. Indeed, the investigations have been continued for years and the summary of studies of ordered silica films grown under various experimental conditions (metal support, partial pressure, annealing temperature, doping) is presented in the form of a minireview in Ref. W1.

3.2 Crystalline silica sheet grown on the Ru(0001) surface*

An ultrathin silica film grown on the Ru(0001) surface was synthesized at the Fritz Haber Institute and analyzed using the scanning tunnelling microscopy (STM), the infrared reflection absorption spectroscopy (IRAS), the x-ray photoelectron spectroscopy (XPS), the near-edge x-ray absorption fine structure (NEXAFS) spectroscopy and the low energy electron diffraction (LEED) techniques. Results of these experiments are presented in detail in Refs. W2 and W3. Here, I report the most important observations, such as: i) the film is a crystalline material; ii) the building units are arranged in a hexagonal pattern (Figure 3.2a); iii) the pattern has a 550 pm lattice constant; iv) the IR spectrum reveals a very intense and sharp absorption peak at around 1300 cm^{-1} (Figure 3.2b).

The structure of the ultrathin silica layer on the Ru(0001) substrate was determined applying the educated guess approach using agreement with the experimental data as the fitness measure. The resulting structure consists of corner-sharing [SiO_4] tetrahedra arranged in a flat honeycomb-like pattern, similarly to the silica structure adapted on the Mo(112) surface.[7] The double-layered form of this film (Figure 3.2c) is the particular feature differentiating it from the structure adapted on the Mo(112) support (Figure 3.1a). Since this structure can be described as two mirror images connected via Si-O-Si bridges aligned perpendicularly to the layers, it is referred to as a double-layer (or bilayer) structure, as opposed to the monolayer structure formed on the Mo(112) support.

This atomic structure model was tested against all experimental data, in particular the experimental IRA spectrum (Figure 3.2b), leading to almost perfect agreement. The description of the applied educated guess procedure and comparison of the experimental data

* Based on Refs. W2 and W3

with the computationally simulated properties of the structure model are given in more detail in Refs. W2 and W3.

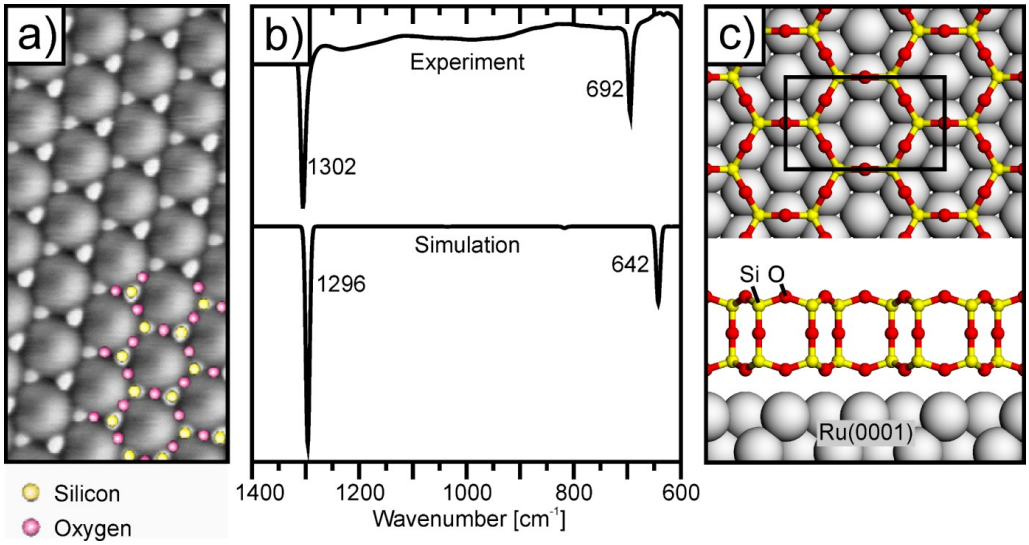

Figure 3.2: a) STM image of silica film grown on the Ru(0001) surface, partially superimposed with the top view of the models shown in c); b) experimental IRA spectrum of O-rich silica film grown on Ru(0001) compared with spectra simulated for the model presented in c); c) Top and side views of the ‘‘O-poor’’ model of silica film on Ru(0001), the orthogonal unit cell is indicated. a) and c) are reprints from Ref. W2, b) from Ref. W3.

3.3 Electronic structure of ultrathin crystalline silica films*

Further experimental investigations of the ultrathin crystalline silica film grown on the Ru(0001) surface demonstrated that the excitation energies related to the Si2p and O1s XPS signals can be tuned using additional ultra high vacuum (UHV) annealing cycles. This XP spectrum alteration process is fully reversible upon the sample reoxidation. On the other hand, virtually no change is observed in the IRA spectra. These and other experimental results clearly indicate that the chemical composition of the film is not affected and that the excitation energies change rather due to adsorption/desorption of the interfacial oxygen atoms, *i.e.* the oxygen atoms adsorbed directly on the metal surface and not bound to the silica layer.

* Based on Ref. W3

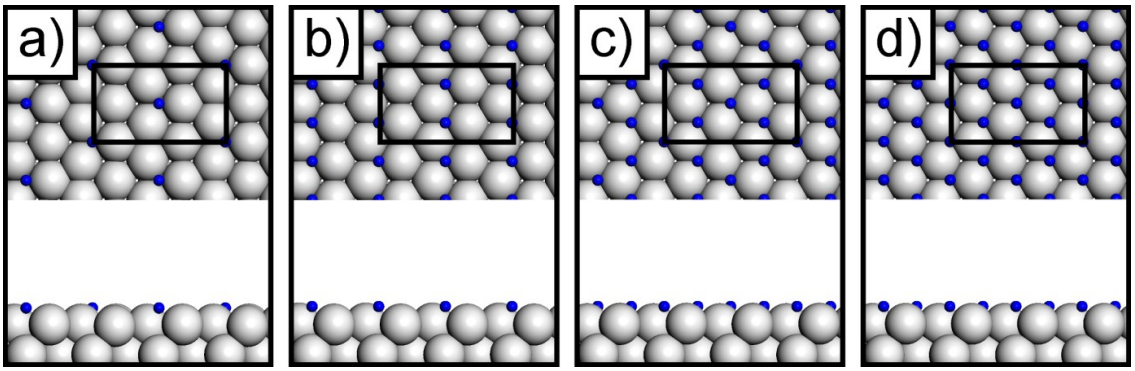

Figure 3.3: Atomic models of oxygen structures adapted on the Ru(0001) surface used in the educated guess search: a) 2O/Ru;[43] b) 4O/Ru;[44] c) 6O/Ru;[45] d) 8O/Ru.[46] The black outline indicates the orthogonal (2 × 2) surface unit cell. O, blue; Ru, gray. Adapted from Ref. W4.

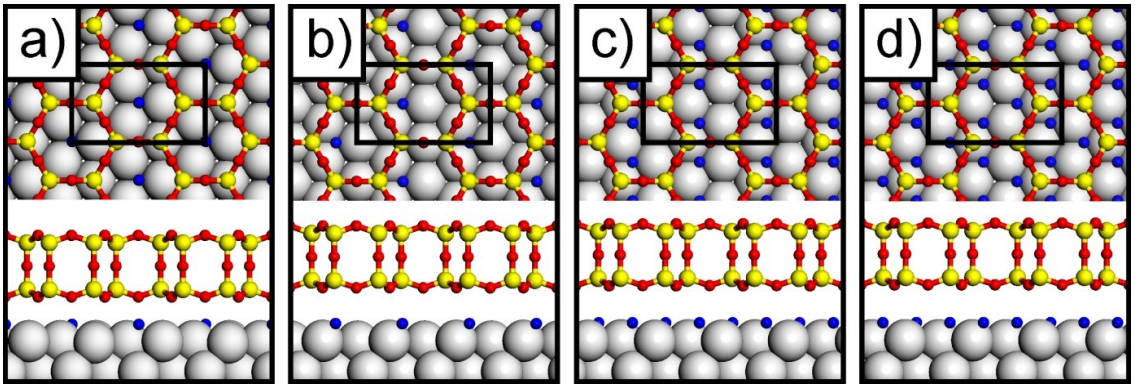

Figure 3.4: The most stable structure models (Si, yellow; O, red or blue; Ru, gray) of: a) $(SiO_2)_8 \cdot 2O/Ru$; b) $(SiO_2)_8 \cdot 4O/Ru$; c) $(SiO_2)_8 \cdot 6O/Ru$ and d) $(SiO_2)_8 \cdot 8O/Ru$ structure models. The black outline indicates the orthogonal (2 × 2) surface unit cell. Adapted from Ref. W3.

In order to check whether the observations can be explained with the change in arrangement of the interfacial oxygen atoms occurring during the UHV procedure, I performed an educated guess search for possible oxygen surface structures. Using the periodic structures adapted by oxygen atoms adsorbed on the pristine Ru(0001) surface[43-46] (Figure 3.3) and the model of silica double-layer (Figure 3.2c), I constructed new surface structure models that contained one to four O atoms adsorbed per the *p*(2 × 2) Ru(0001) surface unit cell, i.e. two to eight per the orthogonal (2 × 2) Ru(0001) surface unit cell. The most stable models are presented in Figure 3.4. While the IR spectra simulated for all these models are virtually identical (Figure 3.5), the simulated O1s XP spectra clearly indicate change in the electronic state of the oxygen atoms (Figure 3.6). In order to check whether the thermodynamic stability of the species containing different amount of interfacial oxygen per the unit cell depends on temperature, I constructed the Gibbs free energy diagram (Figure 3.6c). Indeed, the plot

demonstrates that as the temperature increases the thermodynamic preference shifts towards the structures that contain smaller amount of interfacial oxygen.

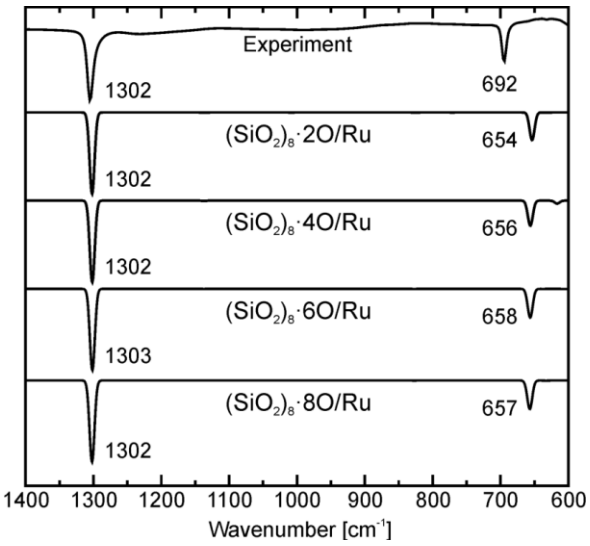

Figure 3.5: Experimental IRA spectrum of the silica film grown on Ru(0001) compared with the spectra simulated for the model systems presented in Figure 3.4. Reprint from Ref. W3.

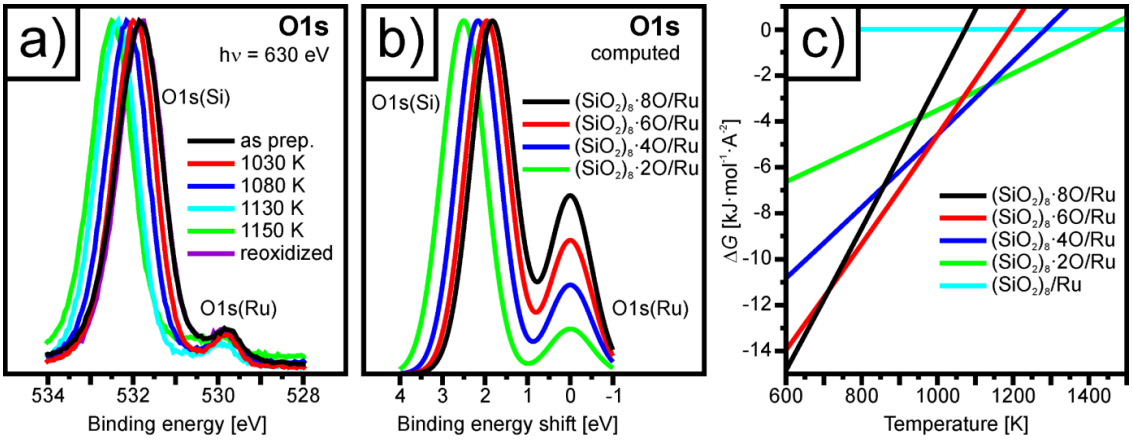

Figure 3.6: a) The O1s (hν = 630 eV) region in the XP spectra of SiO_2/Ru(0001) as a function of temperature of the UHV annealing procedure and after sample reoxidation at 1140 K for 10 min in 2 × 10^{-6} mbar O_2; b) simulated shape of the XP spectra for the most stable double-layer $(SiO_2)_8$·*n*O/Ru models (Figure 3.4), the shift of the binding energy is given with respect to the O1s(Ru) peak, attenuation is not included; c) the Gibbs free energy of oxygen adsorption $\Delta G(T, p = 2 \times 10^{-6}$ mbar) as a function of temperature. O1s(Ru/Si) denote signals assigned to oxygen atoms bonded with the indicated atomic species. Figure adapted from Ref. W3.

These theoretical results confirm that adsorption/desorption of the interfacial oxygen atoms triggered by the oxidation/UHV annealing processes is a plausible reason for the experimentally observed changes in the XP spectra. This topic is covered in more detail in Ref. W3.

3.4 Monolayer, bilayer, and three-dimensional networks*

The crystalline ultrathin silica film formed on the Ru(0001) surface had been grown exclusively in the double-layer form.[W2, W3, W5] Further experimental investigations revealed another phase of the ultrathin crystalline film formed under "Si-poor" conditions (Figure 3.7a). The lattice constant of this film matches the periodicity of the double-layer film (~540 pm). On the other hand, the XP spectrum (Figure 3.7b) is qualitatively different from that obtained for the bilayer phase (Figure 3.6a). Also the IRA spectrum (Figure 3.7c) reveals different characteristics (*cf*. Figure 3.5) but is noticeably similar to that of the silica film formed on the Mo(112) support (Figure 3.1b).

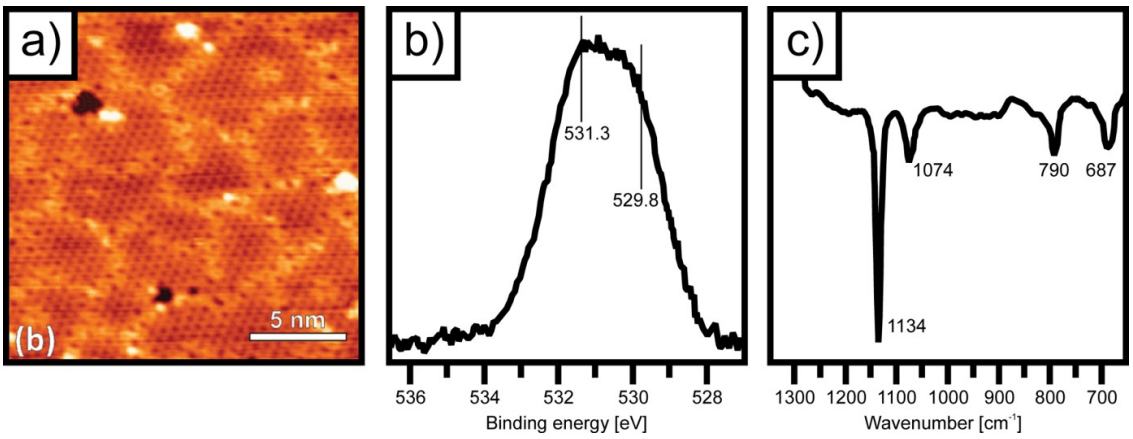

Figure 3.7: Experimental data acquired for the ultrathin silica film formed on the Ru(0001) support under "Si-poor" conditions: a) STM image; b) XP-spectrum (the two indicated binding energy values are obtained after the deconvolution procedure); c) IRA-spectrum. Adapted from Ref. W4.

In order to determine the structure adapted by silica under "Si-poor" conditions, I applied the educated guess approach. Due to the similarity between the IRA spectrum obtained for this film (Figure 3.7c) and the spectrum reported for the monolayer silica formed on the Mo(112) support (Figure 3.1b), I employed exactly that monolayer structure (Figure 3.1a) as the silica ad-layer which is deposited on the Ru(0001) support. Since for the double-layer films the number of the interfacial oxygen atoms can be varied (Section 3.3), I considered such a possibility for the monolayer film model as well. Thus, I included monolayer structures that contain zero to five interfacial oxygen atoms per the primitive (2 × 2) Ru(0001) surface unit cell, *i.e*. zero to ten per the orthogonal (2 × 2) Ru(0001) surface unit cell.

* Based on Ref. W4

In order to investigate the stability of various silica phases as the function of the applied experimental conditions, I created a two-dimensional phase diagram that includes all the monolayer and double-layer silica structure models described in this and previous Sections. Also the pristine ordered O-Ru(0001) structures (Figure 3.3) were included. The resulting phase diagram is presented in Figure 3.8.

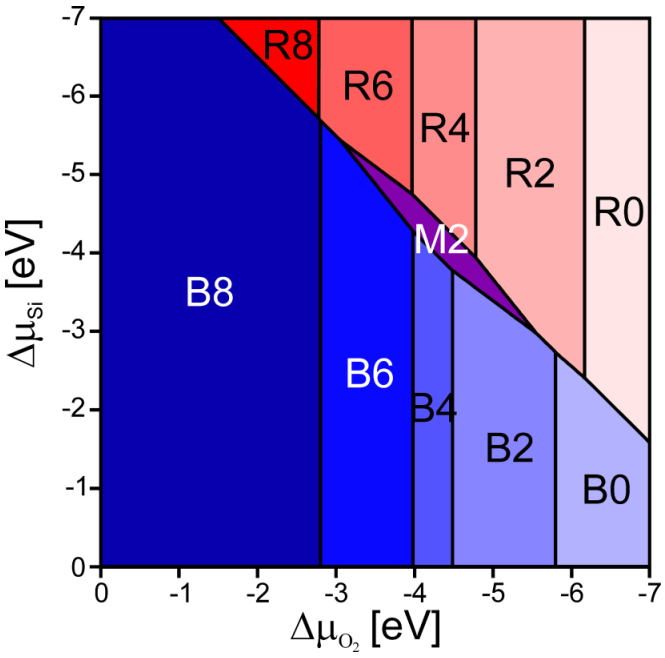

Figure 3.8: Calculated phase diagram of 2D crystalline silica on the Ru(0001) substrate as a function of relative chemical potentials $\Delta\boldsymbol{\mu}_{\mathbf{Si}}$ and $\Delta\boldsymbol{\mu}_{\mathbf{O_2}}$. B, M, and R letters stand for bilayer, monolayer, and oxygen-covered ruthenium regions of stability, respectively. Numbers denote the amount of oxygen atoms in the orthogonal (2 × 2) surface unit cell. Adapted from Ref. W4.

The phase diagram reveals a narrow region of chemical potentials in which a monolayer structure is the thermodynamically preferred structure. This region, denoted in Figure 3.8 as M2, is situated at the border between the bilayer species (B0-B8) and the oxygen-covered Ru(0001) surface species (R0-R8). This sole monolayer model visible on the phase diagram is presented in Figure 3.9a. XP and IRA spectra simulated for this model (Figure 3.9b,c) yield almost perfect agreement with the experimental data obtained for the ultrathin silica film formed on the Ru(0001) support under "Si-poor" conditions (Figure 3.7b,c), thus validating this structure model. The relatively small size of the M2 area indicates that the corresponding structure can be obtained experimentally only in a narrow region of conditions, *i.e.* temperature and partial pressures. This is in line with the fact that initially the crystalline ultrathin silica film grown on the Ru(0001) surface was found in the double-layer structure and the monolayer phase was obtained only after an extensive search.

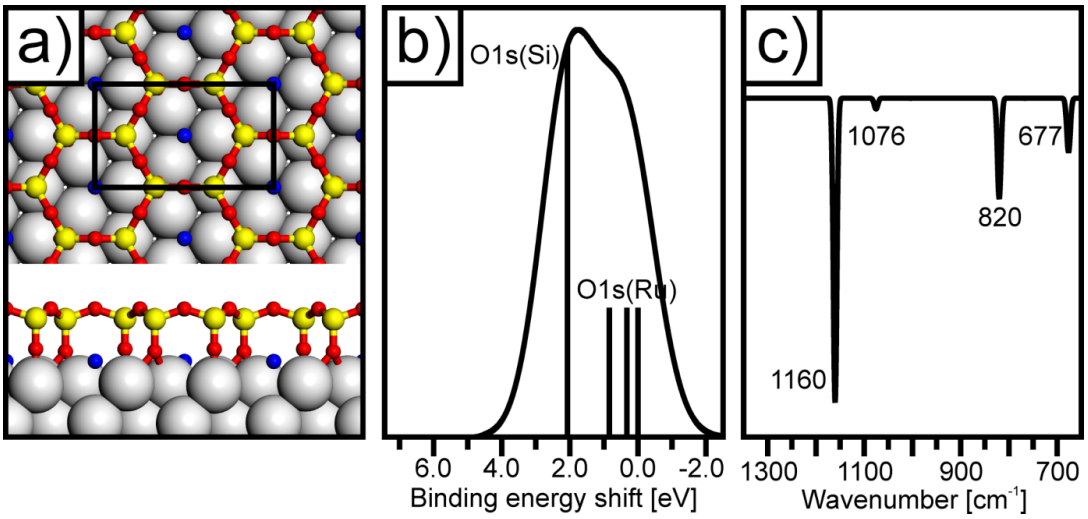

Figure 3.9: The only monolayer model visible on the calculated phase diagram (Figure 3.8): a) the side and top views; Si, yellow; O, red and blue; Ru, gray; b) and c) are simulated XP and IRA spectra, respectively; the bar height in b) is proportional to the number of the respective O atoms in the structure; the BE shifts are given with respect to the lowest O1s state; O1s(Ru/Si) denote signals assigned to oxygen atoms bonded with the indicated atomic species. Adapted from Ref. W4.

Detailed information on applied experimental conditions and used computational methods as well as the entire discussion is presented in Ref. W4.

3.5 Atomic structure of ultrathin Fe-silicate film*

In previous Sections, I briefly reviewed joint theoretical and experimental investigations of structure and properties of monolayer and double-layer SiO_x films formed on the Ru(0001) support. Further experimental investigations revealed that it is also possible to obtain Al-doped ultrathin silica films that retain the double-layer structure and which can be used as "model systems for surface-science studies of the inner walls of zeolite pores."[W6] Applying a similar doping technique, ultrathin Fe-silicate films with varying iron content were prepared as well.[W7]

Since a joint theoretical and experimental study determined the monolayer silica structure,[7] the experiments were supplying unambiguous data that could guide the computational effort in all projects described in previous Sections. However, in the case of the Fe-silicate films, the measured data were initially misinterpreted. As a consequence, it was the computational study

* Based on Ref. W7

that led to determination of the structure and explanation of the properties of that film. In this Section, I briefly report the course and the results of this combined experimental and theoretical investigation.

STM image and IRA spectra of the selected Fe-silicate films are presented in Figure 3.10. The STM image of the $Fe_{0.2}Si_{0.8}O_2$ film (Figure 3.10a) reveals two crystalline phases separated with a boundary. Both phases share the hexagonal honeycomb-like structure with the lattice constants of 550 pm and 525 pm, respectively. The latter phase reveals in addition a moiré pattern with ~220 pm periodicity. The LEED image displayed in the inset of Figure 3.10a represents a superposition of two (2 × 2) diffraction patterns rotated with respect to each other. This corresponds to the presence of two phases whose unit cells are oriented with respect to each other by 30°. The moiré pattern observed in the STM experiment is also reflected in the LEED image (the integer spots for one of the unit cells are surrounded by the satellite spots). Additional experiments show that the moiré-patterned area increases linearly with the amount of iron in the system.

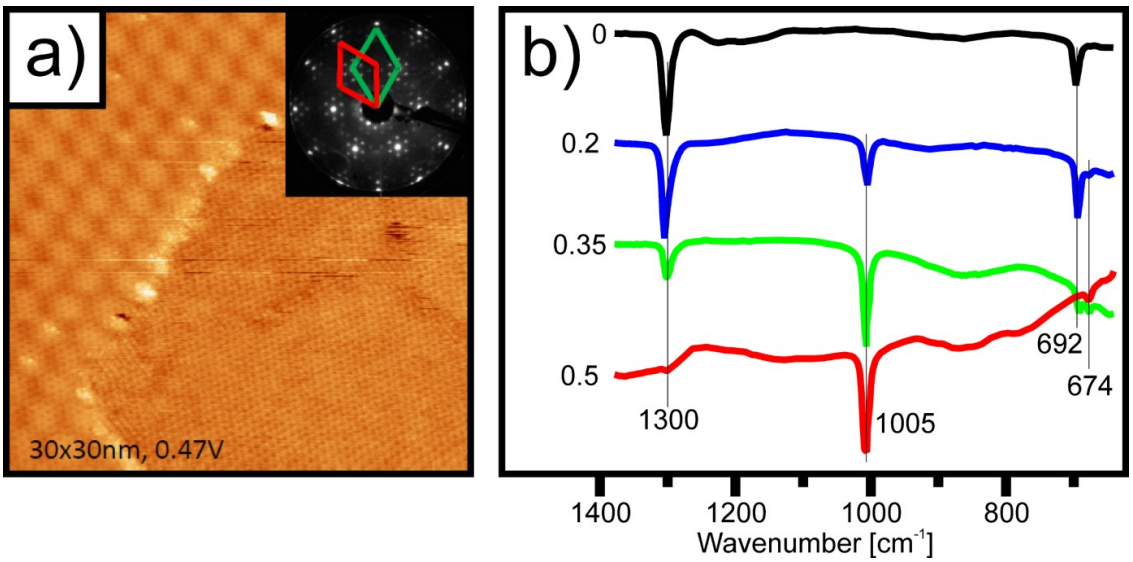

Figure 3.10: Experimental data acquired for the ultrathin Fe-silicate film formed on the Ru(0001) support: a) the STM image of the $Fe_{0.2}Si_{0.8}O_2$ film and the LEED image with two indicated unit cells (in the inset); b) IRA-spectra of the $Fe_xSi_{1-x}O_2$ films as a function of the Fe content (x) as indicated. Adapted from Ref. W7.

A series of IRA spectra obtained for various concentrations of Fe-doping (x) is shown in Figure 3.10b. The spectrum for $x = 0$, *i.e.* for the pure silica double-layer, is identical to that presented in previous Sections (*e.g.* Figure 3.5) and contains two peaks, one at 1300 cm^{-1} and another one at 692 cm^{-1}. The spectrum for $x = 0.2$, *i.e.* for the film with 1:4 iron to silicon ratio, contains additional features. The most apparent one at 1005 cm^{-1} and nearly

unnoticeable peak at 674 cm^{-1}. As the relative amount of iron in the film increases (x = 0.35), the latter two peaks gain intensity while the signals present for the pristine silica become less pronounced. Ultimately, for x = 0.5 the 1300 cm^{-1} and 692 cm^{-1} peaks disappear and only the 1005 cm^{-1} and 674 cm^{-1} bands remain. The changes in composition affect only the intensities but not the positions of the discussed four bands. The exactly opposite behaviour is observed for the Al-silicate films for which Al-doping changes the positions of the bands but not their intensities.[W6]

In order to computationally determine the structure of the Fe-silica ultrathin film formed on the Ru(0001) surface, an educated guess was used. Since the absorption peaks related most probably to the double-layer silica are experimentally observed even for relatively high iron content (x = 0.35 in Figure 3.10), the double-layer models of pristine silica film discussed in the previous Sections (*cf*. Figure 3.2c and Figure 3.4) were natural starting points for the investigation. The experiment was performed under the conditions similar to those used for the double-layer sample reoxidation (Section 3.3), *i.e*. in a relatively oxygen-rich atmosphere. Since the atomic structure of the iron-containing ad-layer was unknown, in particular the amount of oxygen species, I considered one of the models containing interfacial oxygen shown in Figure 3.4. On the other hand, I also considered the possibility of an extensive silica layer reconstruction resulting in formation of covalent ruthenium-adlayer bonds. Thus, in order to leave the metal surface accessible for formation of such bonds, I finally employed the model with small density of the interfacial oxygen, *i.e*. the $(SiO_2)_8$·2O/Ru double-layer model (Figure 3.4a).

The Fe-silicate structure models were created in an iterative manner in order of increasing iron content. In the first step, I created eight structure models of the $FeSi_7O_{16}$·2O/Ru(0001) composition. These models were prepared by substitution of one of eight silicon atoms (each time a different one) with an iron atom and subsequent optimization. The most stable of these structures was selected as a template (seed) for the next iteration. This template was used to generate the $Fe_2Si_6O_{16}$·2O/Ru(0001) models by substituting one of seven remaining silicon atoms. The resulting models were optimized and the most stable one was selected as the template for the $Fe_3Si_5O_{16}$·2O/Ru(0001) structure. This scheme was repeated until the $Fe_4Si_6O_{16}$·2O/Ru(0001) models were generated and optimized.

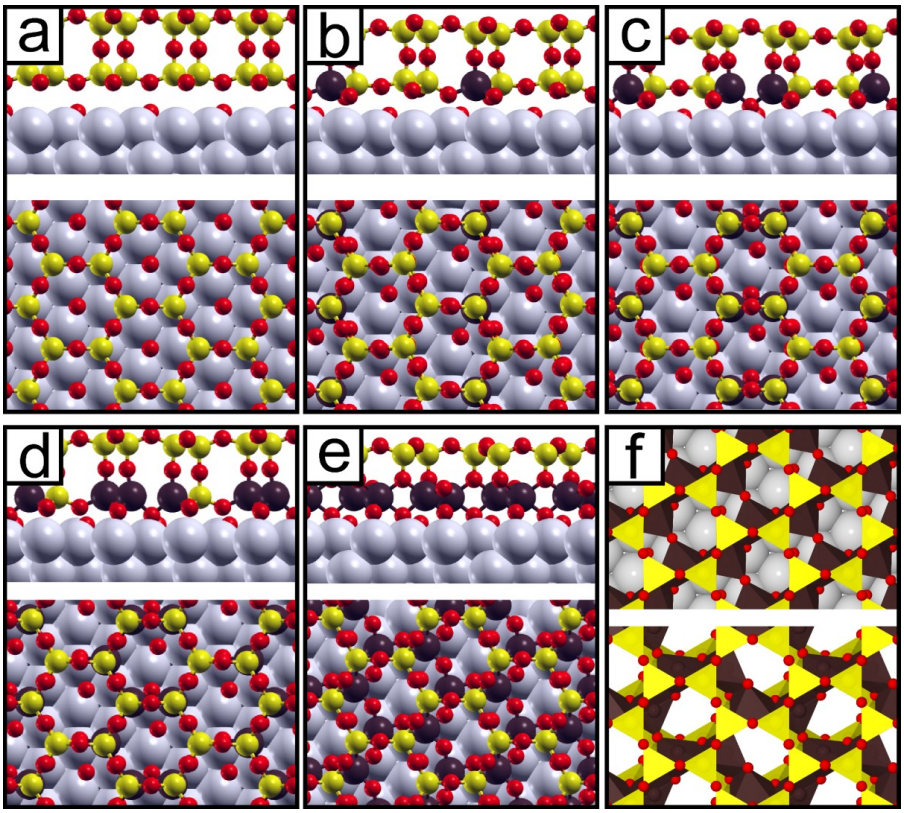

Figure 3.11: Top and side views of the most stable structures found by DFT at various Fe/Si ratios: (a) Si_8O_{16}·2O/Ru(0001); (b) $FeSi_7O_{16}$·2O/Ru(0001); (c) $Fe_2Si_6O_{16}$·2O/Ru(0001); (d) $Fe_3Si_5O_{16}$·2O/Ru(0001); (e) $Fe_4Si_4O_{16}$·2O/Ru(0001); (f) top views of $Fe_4Si_4O_{16}$·2O/Ru(0001) (top) and dehydroxylated nontronite[47] (bottom) in polyhedral representation. Si, yellow; Fe, dark violet; O, red; Ru, gray. Reprint from Ref. W7.

The most stable structures of each investigated composition, *i.e.* $Fe_nSi_{8-n}O_{16}$·2O/Ru(0001) with n = 0 - 4 are shown in Figure 3.11. The first of the models presented there (n = 0) is the pure silica template structure identical to the one presented in Figure 3.4a. The models presented in Figure 3.11b-d (n = 1 - 3) maintain the double-layer structure of the pure silica template with Si-O-Si and Fe-O-Si linkages perpendicular to the metal substrate, although some of the oxygen atoms of the bottom layer are displaced towards the metal. Energetically the most preferable distribution pattern for the iron atoms is the nearest-neighbour arrangement. Unexpectedly, the most stable structure of the $Fe_4Si_4O_{16}$·2O/Ru(0001) composition (Figure 3.11e) undergoes a reconstruction during which the linear Fe-O-Si linkages are converted to planar (Fe-)$_2$O-Si connections and the similarity to the silica bilayer model (Figure 3.11a) is lost. However, the topmost half (the silica layer) retains the silica monolayer structure (*cf.* Figure 3.9a). In fact, the most stable $Fe_4Si_4O_{16}$·2O/Ru(0001) structure can be viewed as a layered film with silica monolayer formed on top of an ultrathin iron oxide film. This resembles the layered structure of nontronite, a natural clay mineral containing iron (*cf.* Figure 3.11f).

In order to assess the relative energetic stability of models with different content of iron, I calculated the energies for reactions that follow the equation:

$$n\mathrm{FeSi_7O_{16}} \cdot 2\mathrm{O/Ru(0001)} \rightarrow \mathrm{Fe}_n\mathrm{Si}_{8-n}\mathrm{O_{16}} \cdot 2\mathrm{O/Ru(0001)} + (n-1)\mathrm{Si_8O_{16}} \cdot 2\mathrm{O/Ru(0001)} \quad (3.5.1)$$

where n = 1 - 4 accounts for the amount of iron in the system. The respective energies given in Table 3.1 show that energetically the least and the most stable of the tested structure models are $Fe_3Si_5O_{16}$·2O/Ru(0001) and $Fe_4Si_4O_{16}$·2O/Ru(0001) models, respectively. The reaction given in Eq. 3.5.1 describes a process of segregation of the initially uniform Fe-silica phase ($FeSi_7O_{16}$·2O/Ru(0001)) into the pure silica phase (Si_8O_{16}·2O/Ru(0001)) and a Fe-silica phase with higher iron content; thus a phase separation process. Since the initial compound in Eq. 3.5.1 is always the same, one can directly compare the reaction energies per iron atom (*i.e.* per a unit cell involved) given in Table 3.1. The energy gain resulting from reaction for n = 4 suggests that the Fe-silicate film has a tendency to segregate into the iron-rich and the iron-less domains. In such a case, the IR spectrum of such a film should be a superposition of signals characteristic for both domains. Moreover, the relative intensity of these signals should depend on the iron content.

Table 3.1: The reaction energy ΔE and the reaction energy per iron atom $\Delta E/n$ for the reaction presented in Eq. 3.4.1.

n	ΔE [kJ·mol^{-1}]	$\Delta E/n$ [kJ·mol^{-1}·atom^{-1}]
1	0.0	0.0
2	-1.0	-0.5
3	+66	+22
4	-158	-40

In Figure 3.12, I present IR spectra simulated for all $Fe_nSi_{8-n}O_{16}$·2O/Ru(0001) structure models (Figure 3.11). Not surprisingly, the spectrum simulated for the Si_8O_{16}·2O/Ru(0001) model (top) is in a perfect agreement with the experimental spectrum obtained for the film with zero iron content (x = 0 in Figure 3.10), *i.e.* the pure silica double-layer (the topic of silica double-layer is covered in Section 3.3). On the other hand, spectra simulated for the

$FeSi_7O_{16}$·2O/Ru(0001), $Fe_2Si_6O_{16}$·2O/Ru(0001), and $Fe_3Si_5O_{16}$·2O/Ru(0001) structure models are qualitatively different from all recorded experimental spectra. On the other hand, the spectrum simulated for the $Fe_4Si_4O_{16}$·2O/Ru(0001) model is in a very good agreement with the experimental spectrum obtained for the film with 50% iron content (x = 0.5 in Figure 3.10). Experimental spectra recorded for the films with 20% and 35% iron content (x = 0.2 and x = 0.35 in Figure 3.10, respectively) can be explained as superposition of absorption bands characteristic for the pure silica bilayer phase (x = 0) coexisting with the film with 50% iron content (x = 0.5). This is consistent with the result of the stability analysis.

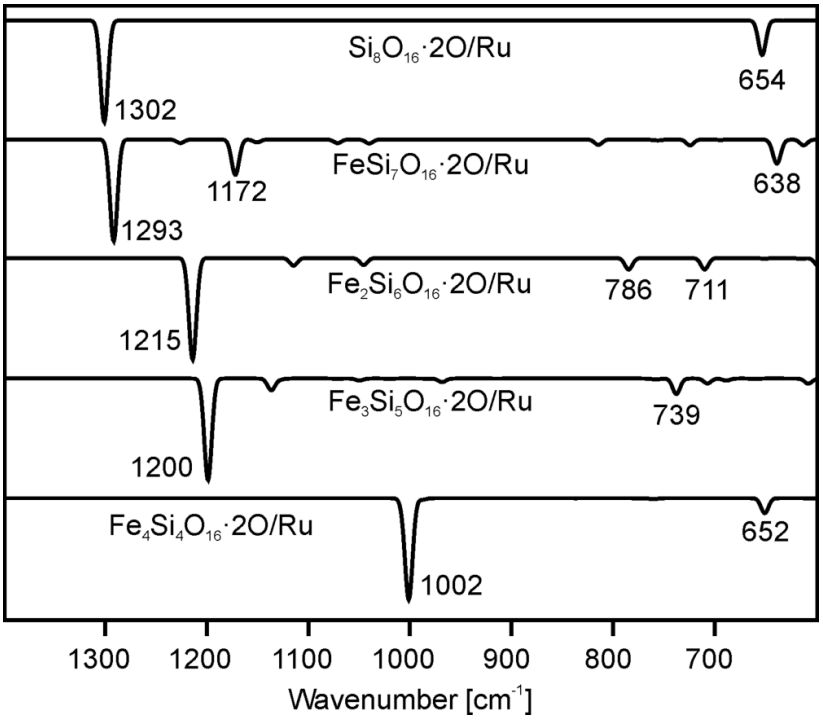

Figure 3.12: IRA spectra simulated for the structural models depicted in Figure 3.11. Reprint from Ref. W7.

Detailed information regarding applied experimental conditions and used computational methods as well as the entire discussion is presented in Ref. W7.

4. The DoDo genetic algorithm package

In 1965, G. E. Moore, one of the future co-founders of the company presently known as Intel Corporation, published a paper on the topic of integrated circuits.[48] In that paper, Moore discussed the state of development of the integrated circuits and perspectives for the future. The author noticed that, due to technological development, the number of components contained in new integrated circuits increases over time while the cost per component decreases. In fact: "The complexity for minimum component costs has increased at a rate of roughly a factor of two per year […]. Certainly over the short term this rate can be expected to continue, if not to increase. Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain nearly constant for at least ten years." Moore's predictions were daring and, as the future showed, also too optimistic. However, the exponential growth in the number of transistors in an integrated circuit over the time is a fact. Thus, the term "Moore's law" was coined and its current version reads: "The number of transistors incorporated in a chip will approximately double every 24 months."[49]

An exponential growth in computational power accompanied by a reduction in price of computers has taken place over the years. The rate of this process can be assessed by comparison of the costs of a hardware with the computational power it offers. A commonly used unit of measure of computer performance is the number of floating-point operations per second (FLOPS). The numbers I present here are cited after Wikipedia.[50] As the first example, I take the Cray X-MP/48 supercomputer produced in 1984 with a theoretical peak performance of 0.8 GFLOPS. Such a supercomputer cost 15,000,000$ so almost 19,000,000$ if the performance is normalized to 1 GFLOPS and about 40,000,000$ after further inflation-adjustment to the 2013 US dollars. As the second example, I chose Sony Playstation 4, a game console of the newest generation released in 2013. The price of a single Sony Playstation 4 unit at that time was 400$ and its peak performance was 1.8 TFLOPS which gives the cost of about 0.2$ per 1 GFLOPS. Thus, a single game console unit (sold in millions) has 2000 times higher computational performance and, at the same time, it is 100.000 times cheaper than the supercomputer available 30 years earlier. In the era of such availability of computational powers, automatization of tedious tasks that require human-computer interface is a matter of time.

The quantum mechanical studies of *e.g.* a particular catalytic site require extensive examination of the potential energy surface. Finding an optimal atomic structure (sometimes

also under specific constraints) is a tedious task. Thus, an efficient method for global exploration of the energy landscape is desired for theoretical chemistry. In order to take advantage of the computational power that is available nowadays, such a method should be automatized to a high degree.

In this Chapter, I present the DoDo package, an implementation of the genetic algorithm for global exploration of the energy landscapes of surface and cluster species. The exploration procedure is fully automatic and, apart from preparation of the input files, no user-program communication is required. Thus, the DoDo program offers the possibility to replace human work by computational effort.

4.1 General package description

The package is subdivided into parts specialized to perform certain actions. The main program: i) communicates with the user; ii) supervises the genetic algorithm run (flowchart in Figure 4.1); iii) stores the population and the history of the run; iv) performs the survival of the fittest algorithm; v) calls specialized genetic operator modules; vi) performs input/output operations; vii) interacts with the queuing system. Since the main program is responsible for communication with the user and with other programs (genetic operators) but is not involved in any computationally demanding operations, it is written in a general-purpose high-level programming language – Python. On the other hand, the operators that employ significant amount of algebraic operations (*e.g.* crossover) are implemented in Fortran, a programming language particularly suited for numeric computation. The local optimization operator represents an exception. Local optimization procedures are commonly used in quantum chemical applications and many programs offering this feature exist. Thus, instead of creating a local optimization code from scratch, I employ external quantum mechanics software packages which offer highly optimized codes that natively run on computational clusters. The input files required by these programs are prepared by DoDo in an automatic manner. Similarly, the output files are analyzed fully automatically. In addition, DoDo communicates with the queuing system. These topics are covered in Section 4.2.2.

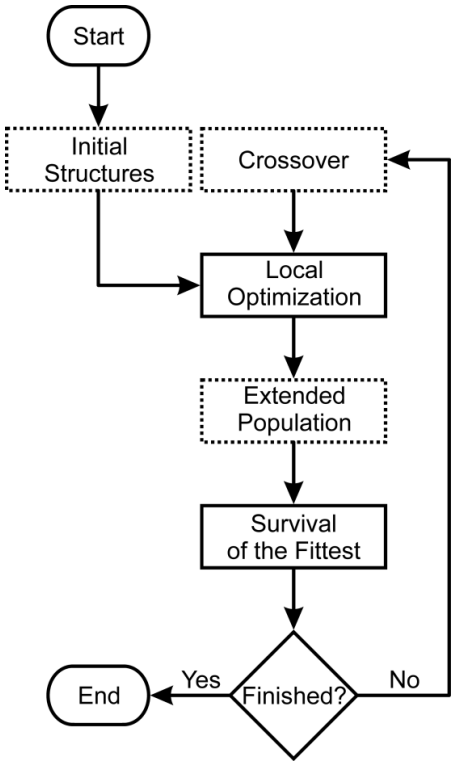

Figure 4.1: Flowchart of the DoDo genetic algorithm package. The procedures that involve similarity recognition are indicated by the dashed lines.

Initially, the functionality of the DoDo package was limited only to surface species. Later, it was extended to cluster species by incorporating an older project, the hybrid *ab initio* genetic algorithm for clusters (HAGA).[51] Since details of HAGA are presented elsewhere,[51, 52] only the implementation of the algorithm for the surface species is presented in the following Sections.

4.2 DoDo implementation

4.2.1 Initial structure models

The initial structure models can be supplied by the user, automatically generated, or obtained by combining both methods. An outline of such a mixed scheme is shown in Figure 4.2. For automatic generation, the user has to specify: i) at least one seed, *i.e*. a surface structure model that can be used as a template; ii) how many initial structures are to be obtained from each seed; iii) desired chemical composition of child structures; iv) atomic radii multipliers.

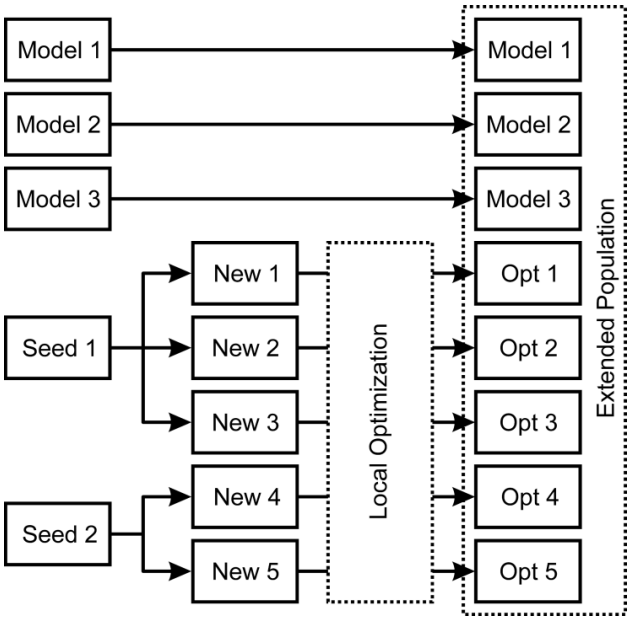

Figure 4.2: Creation of the initial population (represents the first three steps presented in Figure 4.1). Models and seeds represent initial atomistic structures supplied by the user. Models correspond to optimized structures of composition identical to the desired one. Seeds are structures of different composition with respect to the desired one. The structures labelled New 1 – New 5 are obtained from seeds applying the procedure of automatic generation of the initial structures. Opt 1 – Opt 5 denote structures emerging from the New 1 – New 5 models after the local optimization procedure.

The seeds can contain a different number of atoms and even redundant species. The number of atoms to be added to or to be removed from the given seed is computed separately for each of them by assessing the difference from the desired final composition. For example, if the required composition corresponds to a single monolayer of water adsorbed on the MgO(001) surface, Seed 1 in Figure 4.2 can represent the pristine MgO(001) surface (without adsorbed water molecules) while Seed 2 can be a snapshot from a molecular dynamics simulation of solvation of the Na^+–Cl^- pair in four monolayers of water adsorbed on the MgO(001) surface.

Also the number of generated initial structures can be specified for each seed independently. In the example depicted in Figure 4.2, three candidate structures are created starting from Seed 1 and only two from Seed 2.

Seeds can contain “frozen” atoms, *i.e.* the atoms that are not allowed to change their position throughout the genetic algorithm run. This can be used to simulate bulk-like behaviour of parts of the system located farther from the surface.

Automatic generation of a single structure from the given seed is performed in two steps (Figure 4.3). The first step consists of removal of excessive atoms. The second one is addition of missing atoms. Excessive atoms are removed in a loop. The probability of selecting a certain element is proportional to the number of atoms of that kind to be removed (roulette wheel selection). This probability is recalculated whenever the composition of the structure changes, *i.e.* when any atom is removed. When the element is selected, one non-frozen atom of that kind is randomly selected and removed. When deleting one atom leads to disconnection of another atom, *e.g.* removing an oxygen atom from an OH group leaves the hydrogen atom unbound, such a disconnected atom is removed from the structure as well. To evaluate whether an atom is disconnected, distances to its nearest neighbours are calculated. If R_i is the atomic radius of the considered atom, R_j denotes an atomic radius of the jth neighbour, r_{ij} is the distance between these two atoms, and a is the upper atomic radii multiplier given by the user (default a = 1.2), then the atom is flagged as bound if the inequality:

$$r_{ij} < a\left(R_i + R_j\right) \tag{4.2.1}$$

holds for any neighbour j. That is, if the distance to any other atom is smaller than the scaled sum of the atomic radii, the atom is not regarded as disconnected.

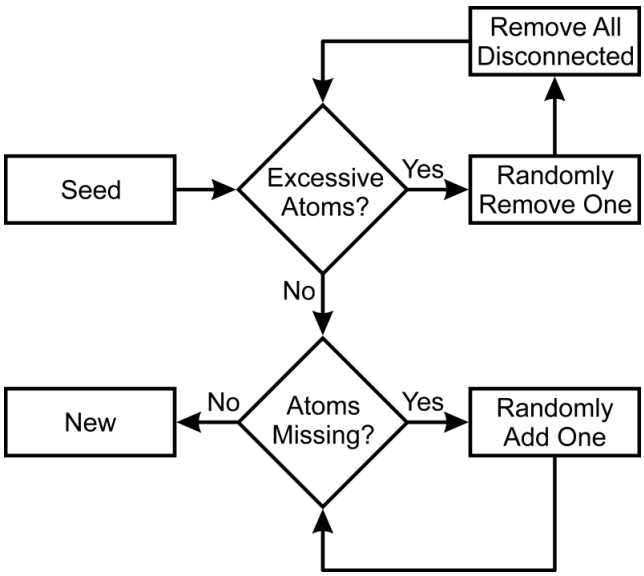

Figure 4.3: Algorithm of automatic generation of a single "New" species from a single "Seed".

The second phase of automatic generation of the initial structure models is addition of missing atoms. Again, atoms are added sequentially applying a roulette wheel selection scheme with

dynamically assessed probabilities. Atoms are distributed randomly on the given surface. New atoms are "adsorbed" on top of the existing surface with possibility to penetrate the surface (Figure 4.4). The depth of penetration is simply the thickness of the non-frozen atomic layer (d_{at} in Figure 4.4) scaled by a factor supplied by user (default 0.3). An atom can become a member of a structure only if it is close enough to be bound to at least one other atom (the check is performed according to Eq. 4.2.1). On the other hand, the candidate atom cannot be too close to any other atom, *i.e.* the atoms are not allowed to extensively overlap. Again, if R_i is the atomic radius of the considered atom, R_j denotes an atomic radius of the jth neighbour, r_{ij} is the distance between these two atoms, and b is the lower atomic radii multiplier given by the user (default $b = 0.7$), then the atom is flagged as overlapping if the inequality:

$$r_{ij} < b\left(R_i + R_j\right) \tag{4.2.2}$$

holds for any neighbour j. That is, if a distance to any other atom is smaller than the scaled sum of the atomic radii, the atom is not allowed into the structure.

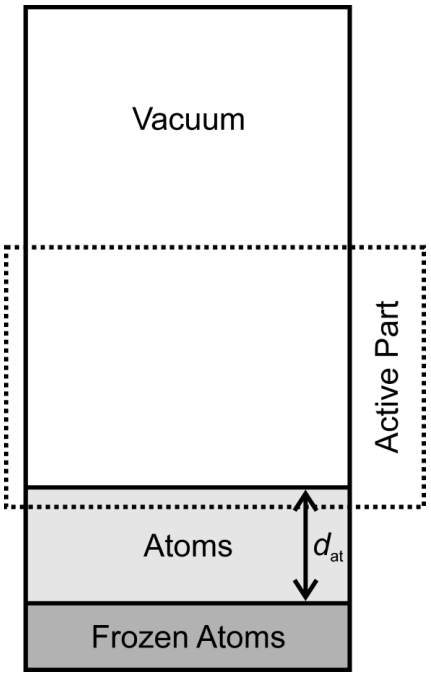

Figure 4.4: Spatial division of the unit cell used during the generation of initial structures. d_{at} denotes thickness of the non-frozen atomic layer.

4.2.2 Local optimization

Following the idea of Judson *et al.*,[23] I employ a Lamarckian (*i.e.* a local optimization) operator. In this implementation, each generated structure is subject to a local optimization procedure. As it was already mentioned in the general introduction to the DoDo package (Section 4.1), the local optimization procedure is not implemented within the DoDo package itself. Instead, I use external quantum mechanics software packages that offer advanced local optimization procedures natively adapted to operate in a multi-core hardware environment. Additionally, the DoDo program is interfaced to the Portable Batch System (PBS), job scheduling software operating on existing massively parallel supercomputers and also on our local clusters.

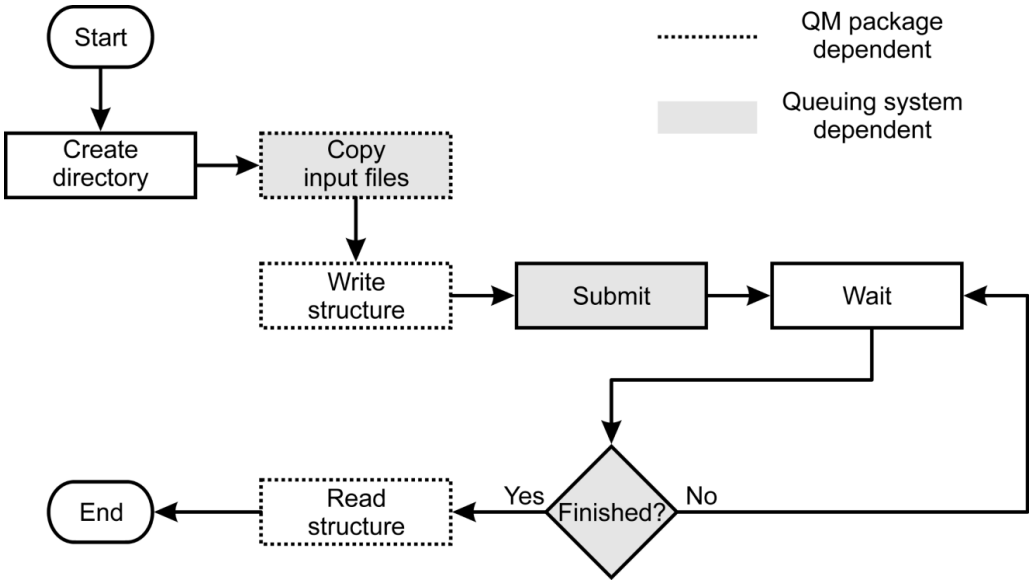

Figure 4.5: Interaction of the DoDo package with the queuing system and the external quantum mechanics package during the local optimization procedure.

Each local optimization is performed in a separate directory and as an independent job according to the flowchart presented in Figure 4.5. The input files required by the queuing system (*e.g.* job script) and the quantum mechanics package (*e.g.* INCAR for VASP) have to be supplied by the user before running the DoDo algorithm. These files are copied into the created working directory. Then, the molecular structure to be optimized is exported according to the formatting required by the QM package (*e.g.* POSCAR for VASP) and the job is submitted to the queuing system using the command compatible with the given job scheduling software (*e.g. qsub* for PBS). The DoDo package stores the unique identifier returned by the queuing system upon the job submission. Job progress is periodically checked

using the identifier and the proper command (*e.g. qstat* for PBS). After the job is finished, the relevant output files created by the quantum mechanics package (*e.g.* OUTCAR for VASP) are parsed in order to obtain the final structure definition and the corresponding energy.

4.2.3 Survival of the fittest

The survival of the fittest algorithm is employed according to the methodology used by Hartke.[30] First, the "extended population" is formed from the optimized structures added to the existing population (Figure 4.1). Creation of the extended population is performed under control of the structural similarity recognition algorithm described below in Section 4.2.5. Then, if the number of elements in the extended population is larger than the given maximum population size, which is usually the case, the extended population is truncated to the maximum allowed size. This is done by sorting the structures according to their energy and by subsequent elimination of the structures with the highest energies. This algorithm ensures that the population consists of the best trial solutions obtained in the course of the global exploration process. This approach was called by Hartke "a strong form of elitism".[30]

4.2.4 Crossover

The first step of the crossover procedure is the fitness evaluation (Figure 4.6). For the fitness assessment, I apply the exponential scheme already used in the HAGA cluster implementation:[51]

$$f_i = e^{-\alpha \varepsilon_i}, \tag{4.2.3}$$

where f_i is the fitness of the ith individual, α is a constant scaling factor, and ε_i is the dynamically scaled relative energy of that individual which takes into account the spread of energies of structures that form the population:

$$\varepsilon_i = \frac{E_i - E_{\min}}{E_{\max} - E_{\min}}. \tag{4.2.4}$$

E_i denotes the energy of an individual and $E_{\max}$ and $E_{\min}$ are the maximum and minimum energies of structures in the whole population, respectively.

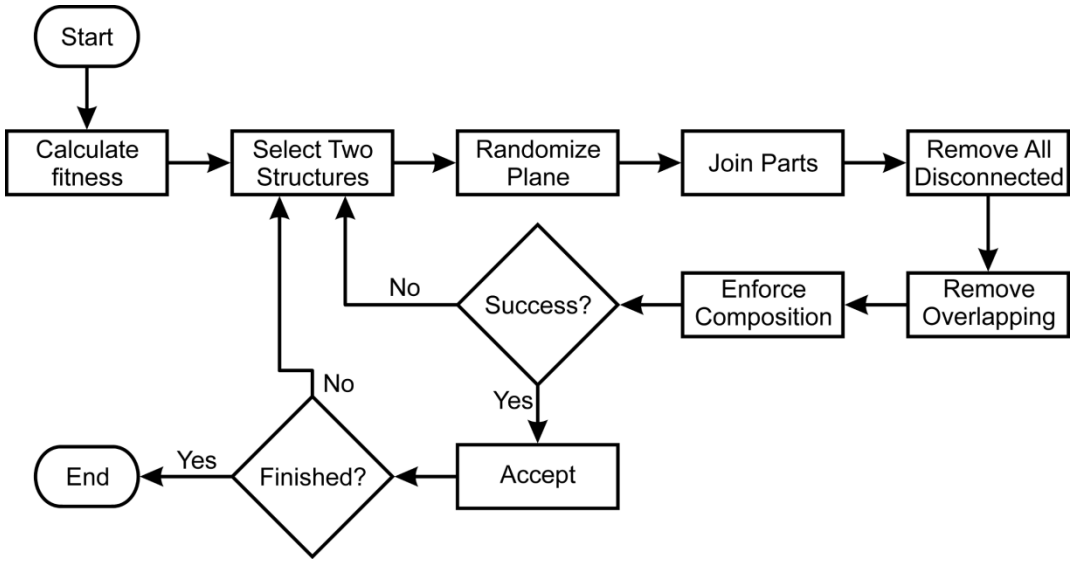

Figure 4.6: Crossover procedure flowchart.

The second step is the selection for mating (Figure 4.6). In this procedure, two elements of the population are chosen to become parents for a new candidate solution. This is performed applying the roulette wheel selection scheme. The probability p_i of selecting an individual *i* is proportional to its fitness:

$$p_i = \frac{f_i}{\sum_i f_i}. \tag{4.2.5}$$

An example of the probability assessment is shown in Table 4.1. In this illustration, the element **A**, characterized by the lowest energy, has the highest probability of selection p_i. In fact, this probability is almost three times larger than that for the element **H**, *i.e.* the energetically least stable element of the population. Although the elements are energetically evenly spaced (1 eV), the probability distribution p_i has the exponential shape, in accordance with Eq. 4.2.3 (*cf.* Figure 2.6).

The mating procedure is performed similarly to the scheme used by Chuang *et al.*[33] The crossover operator acts on structures represented in Cartesian coordinates. Two parent structures are sectioned by an arbitrary plane and the child structure is obtained by combining the complementary parts of the parents (Figure 4.7). However, contrary to the implementation by Chuang *et al.*,[33] the cut is applied to the whole structure, not only to the "active" part. In addition, Chuang *et al.*[33] used only planes perpendicular to the surface while the DoDo genetic algorithm accepts any plane crossing the unit cell.

Table 4.1: Dynamically scaled relative energy ($\boldsymbol{\varepsilon_i}$), fitness ($f_i$) and selection probability (p_i) for the given population of elements calculated using Eqns. (4.2.3-5) with $\boldsymbol{\alpha = 1}$. $\boldsymbol{E_i}$ denotes the energy of an individual.

Element	E_i [eV]	ε_i	f_i	p_i
A	-10.0	0.00	1.00	0.19
B	-9.0	0.14	0.87	0.17
C	-8.0	0.29	0.75	0.15
D	-7.0	0.43	0.65	0.13
E	-6.0	0.57	0.56	0.11
F	-5.0	0.71	0.49	0.10
G	-4.0	0.86	0.42	0.08
H	-3.0	1.00	0.37	0.07

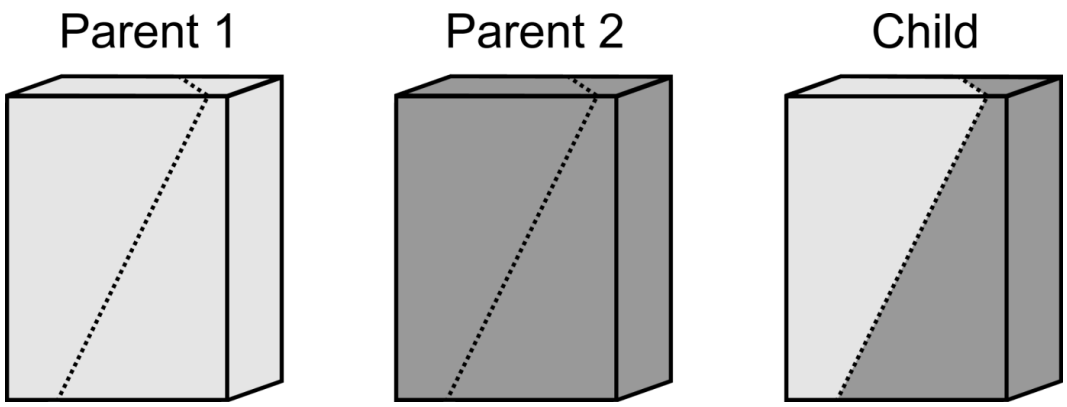

Figure 4.7: Visualization of the mating operation. The cutting plane is indicated by the dashed line.

The obtained child is examined for the presence of non-bound atoms and overlapping atoms. The check is performed as described in Section 4.2.1 (using Eqns. (4.2.1) and (4.2.2)). The non-bound atoms are removed. In the case of overlapping atoms, *i.e*. atoms with too small interatomic distances, the atom(s) to be removed is (are) selected from the involved species in a random manner. Finally, any excessive atoms are removed according to the procedure described in Section 4.2.1.

If a child obtained in such a way is of the required atomic composition, it is accepted to the pool of created children. The algorithm proceeds further according the flowchart presented in

Figure 4.6. That is if the child was not accepted or more children are to be obtained, another cycle of the loop is performed starting from the selection of parents (using already calculated selection probabilities).

4.2.5 Similarity recognition

In order to prevent filling the population with (almost) identical structural models, *i.e.* to maintain structural diversity, a similarity recognition algorithm is implemented. Similarity of two structures is determined by comparison of structural features according to the algorithm presented in Figure 4.8. First, both structures are analyzed to find the atomic species of the least abundance, *e.g.* hydrogen or aluminium for a unit cell with atomic composition of $Ru_{40}Si_7O_{20}AlH$. Then, a grid of cubes encapsulating the whole unit cell is designed. The length of the edge of a cube is selected to match the search radius used for identification of similar atoms. The same grid is designed for both structures.

Then, the algorithm attempts to find a mapping of atoms from one structure to another. Due to the periodic boundary conditions, the procedure of similarity recognition has to comply with the translational symmetry of the structure, *i.e.* arbitrariness in selecting the origin of the unit cell. In order to handle that property of the system, the similarity recognition algorithm attempts to redefine the origin of the unit cell of one of the structures in such a way that atoms belonging to both structures can be superimposed.

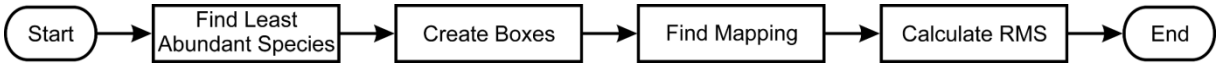

Figure 4.8: Flowchart of the similarity recognition procedure.

Let us consider a pair of atoms P_{AB} created by selecting one atom from structure **A** and one from structure **B**. P_{AB} defines a translation that superimposes these two atoms (Figure 4.9 top). Applying this translation to all atoms that belong to the given structure effectively redefines the origin of the unit cell (Figure 4.9 bottom). If the number of atoms in both structures is equal to N, then N^2 such atom pairs can be created. The structures are flagged as similar if any translation operation leads to such a redefinition of the origin of the unit cell that for all atoms from one structure the corresponding atoms from the other structure are identified and that the relative positions of the corresponding atoms coincide up to a certain threshold. In the example shown in Figure 4.9, structures **A** and **B** are subject to the similarity recognition procedure. Translation applied to the atoms of structure **B** leads to unit cell **B I** and translation

II leads to unit cell **B II**. Although hardly any similarity can be seen for pairs (**A**, **B**) and (**A**, **B I**), structure **B II** is virtually identical to structure **A**. Thus, original structures **A** and **B** in fact represent similar (if not exactly the same) periodic structures.

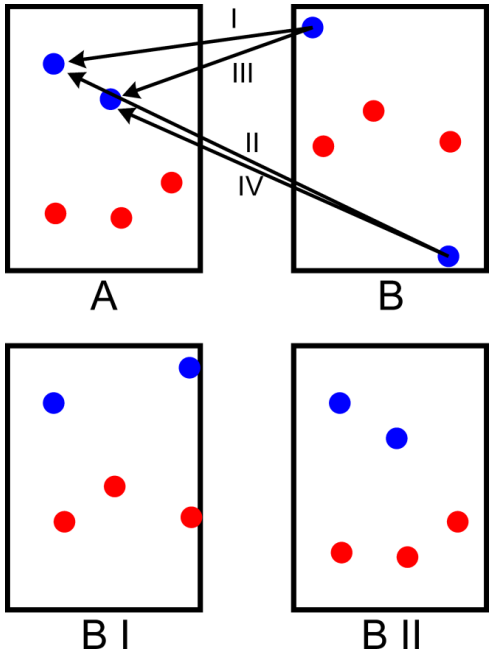

Figure 4.9: Illustration to the procedure of superimposing two atoms belonging to two different structures **A** and **B**. Blue and red dots represent different atomic species. Arrows indicate atomic pairs used to generate translations. **B I** and **B II** are the structures obtained applying respective translation to atoms in structure **B**.

Obviously, superimposing atoms of different species is pointless since these two atoms could not be flagged as corresponding. Thus, only pairs consisting of atoms of the same species have to be examined. Let us consider a single water molecule adsorbed on a 4-layer thick (4×2) MgO(001) surface. This system contains two hydrogen atoms, 33 oxygen atoms, and 32 magnesium atoms per unit cell. Since only translations obtained from the pairs of atoms of the same kind are of interest, this yields $2^2 + 33^2 + 32^2 = 2117$ translations. Here, I remind that in order to flag two structures as similar, a translation operation has to be found that superimposes all atoms of one structure to atoms of the other structure, *i.e.* correspondence for all atoms of all species has to be found. In particular, the necessary condition for similarity of two structures is mapping of the least abundant species, let us call this species Q. In this implementation of the similarity recognition algorithm, only the pairs P_{AB} created by selecting atoms of the Q species are considered since the set of translations defined by these pairs includes all possible combinations of mapping atoms of the Q species between two structures. If it is not possible to obtain full mapping using these translations, the structures are flagged

as different. This approach of focusing on the characteristic atoms of the structure significantly reduces the computational workload. In the considered example, only $2^2 = 4$ instead of 2117 translations would be used.

After a single translation is applied to a structure, each atom from one structure has to be checked for presence of the corresponding atom from the other structure. The brute force (exhaustive) approach is to perform this operation using a nested loop with the outer loop cycling through atoms of one structure and the inner loop cycling through atoms of the other structure. For each such a pair, the difference in positions is checked. If this difference falls within the given threshold, two atoms are flagged as corresponding. Due to the nested loop, the computational cost (complexity) of this process is proportional to the square of the number of atoms ($O(N^2)$).

To reduce this unfavourable scaling, the mapping procedure implemented in the DoDo package employs a uniform grid of cubes ("create boxes" in Figure 4.8). An atom from structure **A** is selected and the cube to which this atom belongs is determined. Since the grid is identical for both structures, there is a corresponding cube placed at the same position in structure **B**. As the length of the edge of the cube matches the search radius (threshold), only that particular cube and its nearest neighbours have to be examined for presence of atoms of the same type as the template atom selected from structure **A**. Thus, for a single atom a constant number of cubes are always checked, *i.e.* the computational cost of the process scales linearly with respect to the number of atoms in structure **A**.

For the given mapping, the root mean square (RMS) of the distances between the corresponding atoms is calculated and compared with the threshold given by the user. If the RMS is smaller than the threshold, the structures are recognized as similar.

The similarity recognition algorithm described above was created in order to maintain structural diversity in the population. Thus, the similarity recognition procedure is applied whenever the population is extended by new structures ("Extended Population" in Figure 4.1). Each new structure is subsequently compared with all members of the population. If the structure is unique (*i.e.* not similar to any member of the population), it is accepted to the extended population. If the structure is similar to a member of population, then the structure with lower energy is accepted to the extended population and the structure with higher energy is discarded. The similarity recognition algorithm is also used whenever a new structure is created ("Initial Structures" and "Crossover" in Figure 4.1). Such a new structure is then

compared with all structures created during the genetic algorithm run (not only with the current population but also with the structures characterized by higher energy and thus rejected in the course of the global exploration process). If the candidate structure is flagged as similar to any other structure, it is immediately discarded. For example, the structure represented in Figure 4.10 by point B would be recognized as similar to the structure represented by point O before entering the local optimization procedure for both shown similarity recognition thresholds.

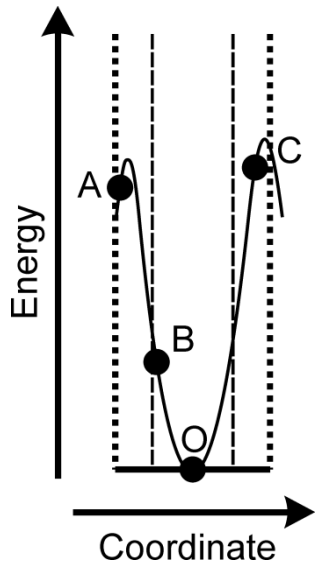

Figure 4.10: Schematic representation of the potential energy surface. A, B, C, and O denote four different points on the PES. Vertical lines encapsulate the area of the PES which represents the configurations recognized as similar to that represented by O using two different similarity recognition thresholds.

The user can influence the extent of structures recognized as similar via the search radius and RMS settings. In the situation presented in Figure 4.10 for the tight similarity recognition criteria (dashed lines), the configuration C, which belongs to the same basin of attraction as the configuration O, would be recognized as similar to the configuration O only after the local optimization procedure. In order to save computational time required for the local optimization procedure applied to the configuration C, the tolerance for deviation of atomic coordinates used in the similarity recognition tool can be increased. This, however, can also lead to improper evaluation of similarity. Such a situation is presented in Figure 4.10 using the dotted vertical lines. There, the configuration A, which belongs to the basin of attraction of a different local minimum, would be mistakenly recognized as similar to the configuration O and discarded from the local optimization procedure. Thus, proper selection of the parameters used for the similarity recognition procedure requires careful testing. The default parameters

are set to allow distinction between two structures of water adsorbed on the MgO(001) surface in a case when the single difference between those two structures is a proton jump between two neighbouring water molecules.

4.2.6 Dynamic steady-state algorithm

In 2003, Hartke investigated structures of neutral water clusters of 2-30 molecules using a genetic algorithm.[53] In that paper, the author reports the following observation: "[…] as soon as the clusters are larger than a minimum size and the potential is not totally trivial, total computation time is dominated by the local optimizations."[53] Our global optimization calculations were performed for periodic systems containing 63-94 atoms per unit cell on a potential energy surface evaluated using a generalized-gradient-approximation-based density functional (Section 5.2). Even though the computational settings used for the energy assessment during the global exploration procedure were already less demanding than the more accurate settings used for the final structure evaluation (*cf.* Sections 5.2.1 and 5.2.2), the total computation time of the process was indeed dominated by the local optimization procedure.

The calculations were performed on a local computer cluster as well as on the North-German Supercomputing Alliance cluster HLRN. At that time, both these installations were controlled by queuing systems which allowed simultaneous execution of multiple jobs in a parallel manner. In addition, each single job was carried out in a multi-core environment using a parallel implementation of the local optimization program. This two-level parallelization significantly reduced the wall-clock time compared to the situation when these jobs were run on a single core. However, despite of that high degree of parallelization, the wall-clock time was still dominated by the local optimization procedures.

This situation is depicted in Figure 4.11a. For simplification, only six children per generation and four permanently available computational slots are considered. Initially, six local optimization jobs are submitted to the queue. All four slots become occupied by a single local optimization and two jobs are scheduled for later execution. After a certain time the optimization process O4 finishes and the slot becomes immediately occupied by the job O5. The genetic algorithm checks the queue status receiving information that one job finished and five other jobs are either running or waiting for execution. Later, the job O2 finishes and the optimization procedure O6 is started. The job O1 finishes but the slot remains idle since no other local optimization procedure is waiting for execution. The queue status is checked again

and the algorithm obtains information that three jobs are finished and three jobs are still running. The jobs O5 and O3 end and the Slots 3 and 4 become idle as well. The queue status is further checked twice. Each time the information that five jobs finished and one is running is passed to the algorithm. Finally, the job O6 finishes and the last slot becomes idle. The queue status is checked again and the algorithm receives information that all six submitted jobs are finished. According to the algorithm presented in Figure 4.1, the optimized structures are added to the population (the similarity recognition procedure is used, *cf*. Section 4.2.5), the survival of the fittest is performed, and the crossover procedure is applied (again using the similarity recognition procedure, *cf*. Section 4.2.5). This leads to creation of six new trial solutions which are submitted to the queuing system as jobs O7-O12. The GA-related operations are not as computationally demanding as the local optimization procedures but nevertheless the time spent on execution of the crossover and similarity recognition algorithms is included in Figure 4.11 and manifests itself as a time shift between checking the queue status and starting the jobs.

In the example presented in Figure 4.11a, *i.e*. the optimization of six children per generation performed on four computational slots, the slots are idle for about 45% of the wall-clock time. The genetic algorithm represented there operates in a generation-based manner, *i.e*. new trial solutions are created only after the whole previous generation of trial solutions was optimized. Thus, the progression to the next generation can be delayed by a single time-demanding local optimization procedure. Such a situation is depicted in Figure 4.11a, where the optimization O6 is the only process occupying any of the computational slots for a relatively long time. In our experience, these hindering optimization procedures were often suffering from convergence problems and were likely to return structure models that were not properly optimized. In such a case, these models, characterized by relatively high potential energy, were anyway rejected during the natural selection procedure (Figure 4.1).

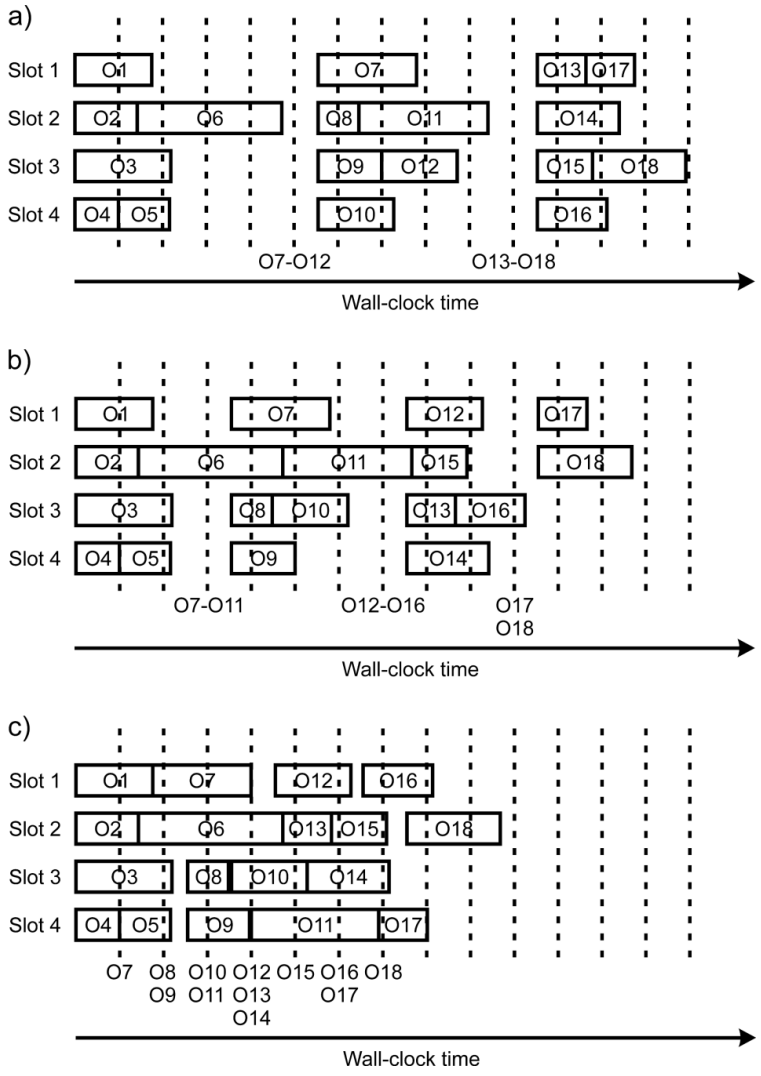

Figure 4.11: Distribution of the computational load as a function of the wall-clock time for three cases of genetic algorithm – queuing system interaction schemes: a) generation-based, b) partially dynamic, and c) fully dynamic. Slots represent available computational resources and rectangles occupation of the particular slot by a local optimization job. The jobs are denoted by O1-O18. Vertical dashed lines represent times in which the queue status is checked and the symbols underneath the lines denote children structures obtained at that time.

In order to reduce the amount of idle time of the computational slots, I modified the original generation-based algorithm in such a way that the progression to the next generation can take place not only when all the scheduled local optimization procedures are finished, but also when one of them has not ended. The distribution of computational load of such a scheme is depicted in Figure 4.11b. This distribution was created applying the same settings as the ones used to obtain Figure 4.11a, *i.e.* four computational slots, a total of 18 optimized structures, and the same computational load used for the corresponding local optimizations. Note that in the first iteration, six children enter the local optimization procedures, but later only five children are created per generation. Similarly to the algorithm discussed above, the number of

the local optimization jobs is checked periodically. However, the number of created children is not constant but selected as either five or six, depending on the number of the local optimization procedures finished since the previous crossover had taken place. For example, the jobs O7-O11 are submitted although the job O6 is still running. Obviously, the structure O6 was not used for creation of the children O7-O11. Similarly, the jobs O12-O16 were submitted before the job O11 finished.

Both these schemes are in fact special cases of a more general algorithm which is based on two parameters: minimum (denoted by a) and maximum (denoted by b) number of optimization jobs present at any time in the queuing system. Whenever the current number of jobs (denoted by c) falls below a, $b - c$ jobs are created and submitted into the queuing system. In the generation-based scheme presented in Figure 4.11a, all jobs are required to finish, thus $a = 1$. Whenever all jobs are finished ($c = 0 < a$), a whole new generation of six ($b - c = 6 - 0 = 6$) children is created and submitted. In the scheme presented in Figure 4.11b, $a = 2$. If all jobs are finished ($c = 0 < a$), a whole new generation of six ($b - c = 6 - 0 = 6$) children is created and submitted. New children are obtained also when one job is still queued or is being executed ($c = 1 < a$) but the number of created children is different ($b - c = 6 - 1 = 5$). The distribution of the computational load for a scheme with $a = b$ is presented in Figure 4.11c. In this scheme, a new child is created immediately after the genetic algorithm notices that any single local optimization finished.

While for the scheme with $a = 1$ (Figure 4.11a) the concept of generation is natural, for $a = 2$ (Figure 4.11b) the generations are still distinguishable, "the number of generations" cannot be used as a unit of the algorithm progress for the scheme with $a = b$ (Figure 4.11c). Instead, I use a more general descriptor, *i.e.* a total number of optimized structures (*cf.* Figure 5.3). For example, all three schemes presented in Figure 4.11 represent runs with 18 locally optimized structures.

5. Application of the DoDo genetic algorithm*

In this Chapter, I present a joint theoretical and experimental study of the interaction of water molecules with the MgO(001) surface. Its aim was to identify the periodic arrangement of water molecules observed experimentally for the monolayer coverage regime. Due to the variety of possible atomic arrangements, *e.g.* the pattern of the hydrogen bonds network or the number of dissociated of water molecules, the DoDo genetic algorithm was employed.

This Chapter is based on the common publication with the experimental group.[W8] Thus, I only briefly summarize the existing experimental and theoretical results, and rather focus on the role of the DoDo program with particular emphasis on the fundamental difference between the techniques of exploration of the potential energy surface applied by our and other theoretical groups. Finally, to validate the employed approach, I present the properties simulated for the most important trial structures and compare them with the experimental results.

5.1 Introduction

The starting point of the theoretical study were the low energy electron diffraction and helium atom scattering investigations of water adsorption on the MgO(001) surface.[54-58] These experiments, performed under UHV conditions applying the temperature programmed desorption technique, clearly show formation of two ordered periodic phases of water that can be observed in the limit of monolayer coverage. The phase characterized by the $c(4 \times 2)$ symmetry was observed for the temperatures of 100-180 K, while the $p(3 \times 2)$ phase, containing a glide plane, was present for 185-221 K. It was shown that the transition from the $c(4 \times 2)$ phase to the $p(3 \times 2)$ one occurs upon heating and is accompanied by partial desorption of water.

An overview of the theoretical studies of the structure of water adsorbed on the MgO(001) surface is given in Ref. W8. Here, I briefly report the methodology in which these structural models were designed, *i.e.* the algorithms that were used to explore the potential energy surface in those theoretical investigations. Ferry *et al.*[59] studied the relative arrangement of rigid water molecules on top of the rigid MgO(001) surface. The search for the most stable

* Based on Ref. W8

conformation of water molecules was performed starting from the (1 × 1) up to the (4 × 2) supercell periodicity. It was found that in the (1 × 1) unit cell the cation site is the preferable adsorption site for a single water molecule. The global exploration algorithm used to obtain trial structures for the larger unit cells is not reported. However, Giordano, Goniakowski, and Suzanne (co-authors of Ref. 59) refer to these calculations as the "semiempirical molecular dynamics simulation",[60] suggesting that the trial configurations were obtained from MD trajectories.

The MD technique was applied also by Odelius[61] in his investigations of the systems with three to six water molecules in the (3 × 2) unit cell. The additional structure models were obtained employing the educated guess approach by simple modification of the models acquired from the MD runs. Cho, Park, and Kim[62] investigated adsorption of one to four and six water molecules in the (3 × 2) unit cell and eight water molecules in the (4 × 2) unit cell. The algorithm used to create models of one to four water molecules adsorbed in the (3 × 2) unit cell is not described, but, in order to include the symmetry features consistent with the experimental findings, the authors applied the educated guess approach to obtain the structure models for six water molecules adsorbed in the (3 × 2) unit cell and eight molecules in the (4 × 2) unit cell.

Delle Site, Alavi, and Lynden-Bell[63] investigated the structure adapted by water molecules adsorbed on the MgO(001) slab. In their models, six water molecules were present per the (3 × 2) surface unit cell. The authors used the structural models proposed in Refs. [59] and [60] as starting structures for the molecular dynamics-based simulated annealing procedures. In the follow-up paper, Lynden-Bell, Delle Site, and Alavi[64] systematically explored the energy landscape using educated guess based grid search approach. The structural models were created to satisfy the observed symmetry constraints, *i.e.* the presence of the glide plane. The search was confined to the system with four intact and two dissociated water molecules. The authors also assumed that each oxygen atom from the water ad-layer is placed on top of the Mg site. Taking these constraints into consideration, ten trial structures were created.

Finally, also Jug, Heidberg, and Bredow[65] investigated structures of water molecules adsorbed on the MgO(001) surface. They used three different surface unit cell models: i) the (3 × 2); ii) the (4 × 2); and iii) the (1 × 1) unit cell. For the first model they used six water molecules per unit cell, for the second model eight and twelve, and for the third model one water molecule per unit cell. All trial structures were obtained using the educated guess approach. In the

following paper,[66] the authors applied the molecular dynamics technique to explore the energy landscape for the systems presented in Ref. 65.

5.2 Case study: Ordered water monolayer on the MgO(001) surface

The results of our joint investigation of the structures of the ordered water monolayer formed on the MgO(001) surface are presented in Ref. W8. In this Section, the original data from Ref. W8 are supplemented with unpublished details and an additional analysis.

Our study of the ordered structure of water adsorbed on the MgO(001) surface included investigation of both experimentally identified translational symmetries, *i.e.* the (3 × 2) and the (4 × 2) surface unit cells. Since the exact number of water molecules contained in a unit cell was not determined, systems with 0.83-1.33 water molecules per single surface cation site were investigated (Table 5.1).

Table 5.1: Water coverages ($n$$H_2O$ molecules per (1 × 1) MgO(001) surface) of the studied systems. Reprint from Ref. W8.

unit cell	$n$$H_2O$ / unit cell	coverage
(3 × 2)	5	0.83
(4 × 2)	7	0.88
(3 × 2)	6	1.00
(4 × 2)	8	1.00
(4 × 2)	9	1.13
(3 × 2)	7	1.17
(4 × 2)	10	1.25
(3 × 2)	8	1.33

The computational study was divided into three phases. In the first phase, a separate investigation of the potential energy surface was performed independently for each of the systems presented in Table 5.1. The aim of the first phase was to obtain a rough estimate of the potential energy surface in order to prescreen trial solutions for the global minimum

structure for each of the investigated systems. The energetically most preferable trial solutions obtained in the first phase were used in the second phase in order to assess the energy ordering of the most relevant structures using more accurate computational settings. At the same time, the second phase served as a verification of the computational settings applied in the first phase. In the case of a severe discrepancy in the relative energies obtained in the first and the second phase, validity of the prescreening would be questionable. The energetically most stable trial solutions obtained in the second phase were selected for the third phase, *i.e.* comparison with the experimental data.

5.2.1 Exploration of energy landscapes

The exploration of the energy landscape was performed using the DoDo genetic algorithm (flowchart: Figure 4.1). For each of the investigated systems, a set of 50-60 initial trial structures was created distributing 2*n*H and *n*O atoms (Table 5.1) according to the algorithm described in Section 4.2.1. The local optimization procedures were performed using the Vienna Ab initio Simulation Package (VASP)[67, 68] along with the Perdew, Burke, and Ernzerhof (PBE)[69, 70] exchange-correlation functional. I employed a 250 eV energy cutoff for the plane wave basis set and a 2 × 1 × 1 Monkhorst-Pack[71] *k*-point mesh for the integrations of the first Brillouin zone. More details are given in Ref. W8.

The procedure of automatic generation of trial structure models is performed in a random manner described in Section 4.2.1. An example of an automatically generated initial trial structure for the $7H_2O/(3\times2)MgO(001)$ system is shown in Figure 5.1. Since the interatomic distances are mostly longer than the thresholds used for bond visualization, hardly any formal bond is indicated in the figure. The consequence of the automated randomized approach is that the quality of the generated trial solutions is not guaranteed to be uniform. Obviously, the randomly generated structures usually belong to different basins of attraction. Thus, successfully applied local optimization procedures lead to structures characterized by distinctive potential energies. In addition, some of the local optimization processes might suffer from convergence problems and thus fail. Combination of these two effects leads to large spread in energies of the structures after optimization procedure as it is shown in Figure 5.2.

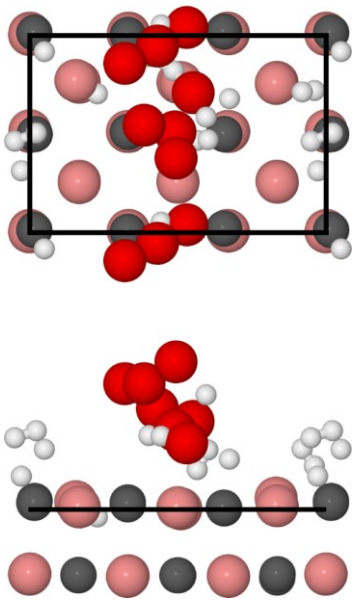

Figure 5.1: Side and top views of an automatically generated initial structure model (Mg gray, O red, H white). The bonds are shown for interatomic distances lower than 110 pm (130 pm for O-O).

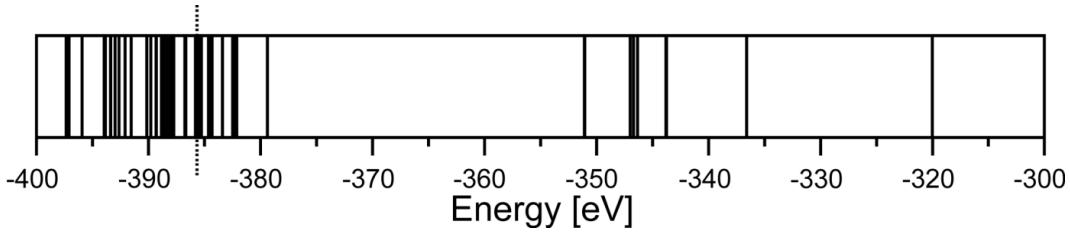

Figure 5.2: Potential energy spectrum of automatically generated trial structure models for the $7H_2O/(3\times2)MgO(001)$ system after local optimization procedure. Each vertical line represents a single model. The dotted vertical line indicates the energy of the least stable structure model accepted in the initial population.

The energy spectrum shown in Figure 5.2 represents the result of the optimization of 50 initial structures generated for the $7H_2O/(3\times2)MgO(001)$ system. The energy ranges from -320.0 eV to -397.4 eV, resulting in a total spread of almost 80 eV but as many as 41 models (80%) fall within the energy difference smaller than 18 eV (20% of the range) from the most stable model. The presence of structures with significantly higher potential energy amongst the initial structure models was the reason to generate more initial trial solutions than there are places in the population. This ensures that not all of the structures fit in the population and the worst (most probably not converged) structures are discarded. In this case, 50 trial solutions were automatically generated, whereas the population size was limited to only 30 structure models. The dotted vertical line in Figure 5.2 indicates the energy of the 30th energetically most stable structure model (-385.7 eV), *i.e.* the last one that entered the population.

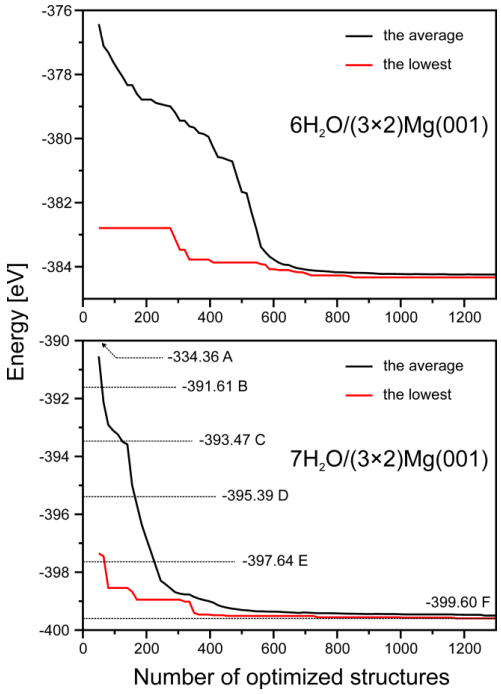

Figure 5.3: The progress of the global exploration of the energy landscape for the $6H_2O/(3\times2)MgO(001)$ and $7H_2O/(3\times2)MgO(001)$ systems. The red line represents the potential energy of the most stable structure found in the course of the optimization process. The black line represents the average energy of all trial solutions present in the population. The energies indicated in the plot for the $7H_2O/(3\times2)MgO(001)$ system correspond to the structures shown in Figure 5.4.

Figure 5.3 demonstrates the progress of the global exploration of the energy landscape for the $6H_2O/(3\times2)MgO(001)$ and $7H_2O/(3\times2)MgO(001)$ systems. The structure models present in the initial population are characterized by relatively large energy spread, which is manifested by a considerable difference between the lowest and the average energy (the left side of the plots). Upon the genetic algorithm run, models with the highest potential energy are gradually replaced with the models of lower potential energy. Already within the next 50 optimized structures (*i.e.* after the total of 100 local optimization procedures), the average energy of the population decreases by about 2 eV for both systems. The average and the lowest energy of the population further decline to finally converge to almost constant values after ~1000 structure optimizations.

The snapshots of structure models obtained during the global exploration are shown in Figure 5.4. The dotted lines in Figure 5.3 indicate the energies of these structures on the energy scale spanned during the genetic algorithm run. Model **A** in Figure 5.4 is characterized by very high potential energy, corresponding to the worst energies present in Figure 5.2. The reason for such a high energy is failure of the local optimization procedure, which results in unphysical

interatomic distances. Interestingly, the structure model already incorporates certain structural features present for lower energy structures, such as six O atoms aligned atop Mg sites forming a plane, one oxygen atom with one hydrogen atom atop pushed out of this plane, and the MgO(001) surface hydroxylated directly below this OH group.

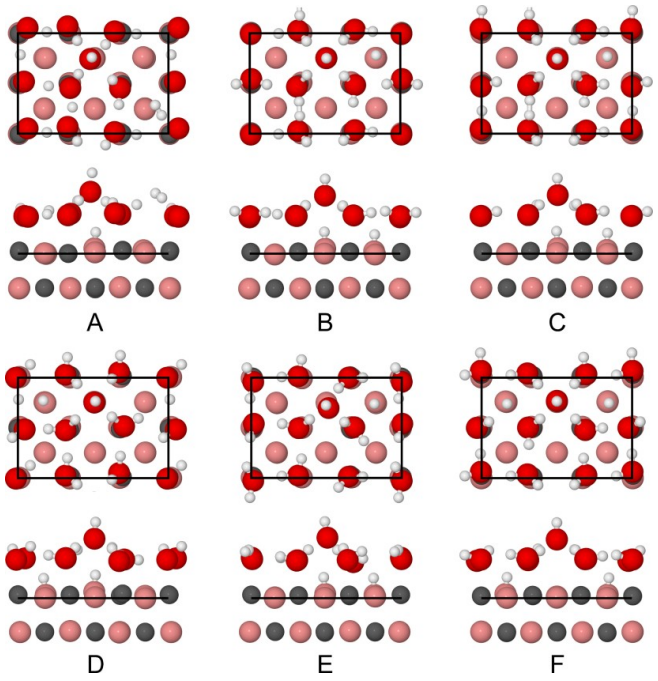

Figure 5.4: Side and top views of the structure models obtained in the course of global optimization for the $7H_2O/(3\times2)MgO(001)$ system (Mg gray, O red, H white). The A-F labels correspond to indications in Figure 5.3. The bonds are shown for interatomic distances lower than 110 pm (130 pm for O-O).

Interatomic distances in model **B** are significantly shorter, which is indicated with formal bonds in pictorial representation. In turn, the potential energy for this model is 57 eV lower than the one for model **A**. The peculiar feature of this model is the H-H bridge connecting two neighbouring water molecules and a single not-saturated O atom in the adlayer.

The potential energy of model **C** is lower by further 2 eV. In this structure, the H-H bridge connecting two neighbouring water molecules is also present, however, all oxygen atoms of the adlayer are already bound to at least one hydrogen atom. The energy of this model corresponds to the average energy of population obtained after ~120 local optimizations. Model **D** contains unbound atoms but in turn the H-H bridge disappeared. The energy corresponds to the average energy of population after ~170 local optimizations.

Model **E** is composed of intact water molecules and OH groups arranged in the hydrogen bond network. The energy of this structure corresponds to the average energy of population

after ~230 local optimizations but is also close to the energy of the most stable structure found after 65 local optimizations. Finally, model **F** is the most stable structure found by the global exploration, which differs from model **E** by arrangement of hydrogen bonds and positions of the surface OH groups.

Changes in energy of the most stable structure for the $6H_2O/(3\times2)MgO(001)$ and $7H_2O/(3\times2)MgO(001)$ systems in the interval of 300 local optimizations is summarized in Table 5.2. In both cases, the most significant energy change occurs within the first in the beginning of the exploration procedure and further global optimization process yields smaller energy changes. For all systems, the global exploration process was ceased after about 1000-1500 local optimization procedures due to only extremely small changes in the lowest and the average energies.

Table 5.2: The energy of the most stable member of population during the global optimization of the $6H_2O/(3\times2)MgO(001)$ and $7H_2O/(3\times2)MgO(001)$ systems taken in 300 optimization intervals. E_{low} and ΔE_{low} denote the lowest energy and the change in the lowest energy with respect to the value from the previous interval, respectively.

Number of optimized structures	$6H_2O/(3\times2)MgO(001)$		$7H_2O/(3\times2)MgO(001)$	
	E_{low} [eV]	ΔE_{low} [eV]	E_{low} [eV]	ΔE_{low} [eV]
50	-382.79	-	-397.35	-
350	-383.78	-0.99	-399.41	-2.06
650	-384.10	-0.32	-399.51	-0.10
950	-384.33	-0.23	-399.56	-0.05
1250	-384.33	0.00	-399.59	-0.03

5.2.2 Refinement of results

The aim of the second part of the project was to refine the structure models obtained during the first phase. The selection of the models was performed using the energy criterion. For each of the investigated systems, the acceptance threshold was set to 30 kJ/mol (0.3 eV) above the energy of the most stable model obtained in the genetic algorithm run. All atomic structure models characterized by the potential energy lower than the threshold were subject

to the final structure optimization and energy evaluation employing a 400 eV energy cutoff for the plane wave basis set and a 4 × 2 × 1 Monkhorst-Pack[71] *k*-point mesh.

The influence of change of the computational settings on the relative energy ordering of the most stable structure models obtained for the $6H_2O$/(3×2)MgO(001) system is shown in Table 5.3. For this particular system, the most stable model found during the GA procedure remains the most stable one also after the refinement procedure. The relative energies of the other models, *i.e.* the energies calculated relatively to the energy of the most stable structure, change by at most 0.015 eV. However, already such a small alteration can be sufficient to change the ordering in a case of nearly degenerated energies (*cf.* local minima **III** and **IV** in Table 5.3).

Table 5.3: Relative energies of the most stable structure models found during the global optimization of the $6H_2O$/(3×2)MgO(001) system before and after refinement.

Local minimum	Relative energy [eV]	
	Global exploration	Refinement
I	0.000	0.000
II	0.058	0.050
III	0.069	0.084
IV	0.089	0.078

For all eight investigated systems, the changes in the energies calculated relatively to the energy of the most stable structure do not exceed 0.05 eV. Also, changes in the atomic structure of the models during the refinement procedure are negligible. Such small energy differences indicate that the lower accuracy settings applied for the global exploration phase only slightly altered the potential energy landscape. Thus, the used settings are validated. The most stable structures identified for all investigated systems after the refinement procedure are given in Ref. W8. None of these most stable structure models, apart from the model for the $6H_2O$/(3×2)MgO(001) system, had been known to the scientific community. The most stable $6H_2O$/(3×2)MgO(001) structure found during our investigations had been proposed by Lynden-Bell, Delle Site, and Alavi.[64]

5.2.3 Comparison of simulated and experimental properties

The third part of the project focused on analysis of the most stable structure models obtained in the second phase. The potential agreement between the experimentally identified properties of the water monolayer and the properties simulated for the structure models would be at least partial validation of the approach used to generate the models, *i.e.* the genetic algorithm.

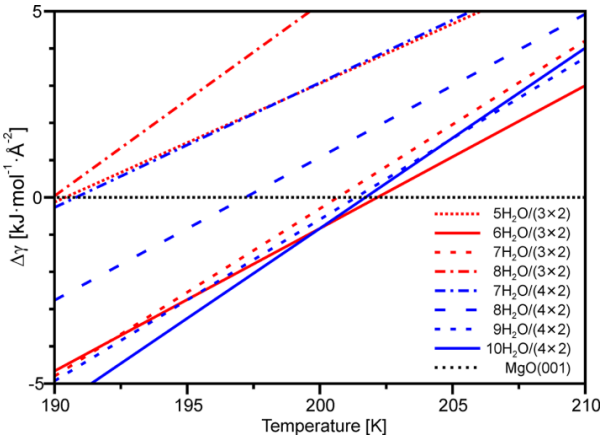

Figure 5.5: Gibbs free energy for adsorption of water molecules on a clean MgO(001) substrate $\Delta\gamma(T,p = 10^{-10}$ mbar) for the most stable nH$_2$O/(3×2) and mH$_2$O/(4×2) structure models as a function of temperature. Reprint from Ref. W8.

In the first step, I compared the relative stability of the most stable structure models obtained for each of the nH$_2$O/(m × 2) MgO(001) systems listed in Table 5.1. Since the structures differ in chemical composition and in surface area per unit cell, this was done using the Gibbs free energy for adsorption of n water molecules on a clean MgO(001) substrate calculated per area according to equations given in Ref. W8. The Gibbs free energies were calculated for a constant pressure $p = 10^{-10}$ mbar, *i.e.* the experimentally relevant value. The result is presented in a graphical form in Figure 5.5. The figure demonstrates that the preference of the most stable surface structure model depends on temperature. The 10H$_2$O/(4×2) MgO(001) structure is stable below 200 K, the 6H$_2$O/(3×2) MgO(001) structure is stable between 200 and 202 K, and above 202 K the most stable model is the pure MgO(001) surface. The water coverage per Mg site equals to: (i) 1.25 molecules for the 10H$_2$O/(4×2) MgO(001) structure; (ii) 1.00 molecules for the 6H$_2$O/(3×2)MgO(001) structure; (iii) 0 molecules for the pure MgO(001) surface. Thus, according to the stability analysis, water molecules should gradually desorb upon the temperature increase. The most stable 10H$_2$O/(4×2) MgO(001) structure model is characterized by the c(4 × 2) symmetry, *i.e.* the atoms are arranged in the centered (4 × 2)

surface lattice. The most stable $6H_2O/(3\times2)$MgO(001) model adapts different symmetry, the $p(3 \times 2)$ symmetry with a glide plane.

The experimental desorption spectra clearly show that water layer gradually desorbs upon the temperature increase,[57] which is consistent with the progression predicted from the stability plot (Figure 5.5). In addition, the experiments show that at low temperature (100-180 K) the water monolayer adapts the $c(4 \times 2)$ symmetry, and for 185-221 K the monolayer is characterized by the $p(3 \times 2)$ symmetry and contains a glide plane.[55] All these symmetry-related features identified in experiments are consistent with properties of the most stable structure models.

The models were further verified by comparison of the vibrational spectra simulated for the $10H_2O/(4\times2)$ MgO(001) and the $6H_2O/(3\times2)$ MgO(001) structures with the experimentally acquired data. Detailed description and discussion that lead to assignment of the observed vibrational modes is given in Ref. W8 and is there summarized as follows: "the excellent agreement of calculated and experimental vibrational spectra strongly suggests that the structural models proposed in the present study realistically represent the water monolayer structures present on the MgO(001) surface." [W8]

5.3 Conclusions

The structural models obtained from the genetic algorithm runs reliably reproduced the experimental data such as adsorption spectra or information about the unit cell symmetry. Moreover, stability analysis describes the temperature dependence of these properties. In principle, these atomic models could have been created using an educated guess approach. This is the case for the $6H_2O/(3\times2)$ MgO(001) system, for which Lynden-Bell, Delle Site, and Alavi already in 2002 designed the model identical to the most stable structure found in this study. However, our investigation not only (i) included extensive examination of the energy landscape, but also (ii) was performed consistently for eight model systems. While the former indicates that the structure models obtained by us are, or at least are close to, global minima, the latter allows us to create a stability plot (Figure 5.5) that reveals which of the investigated systems can be important for the proper description of the temperature-dependent properties of the water monolayer adsorbed on the MgO(001) surface. In this way, we were the first scientific group that identified the significance of the $10H_2O/(4\times2)$ MgO(001) system for the low-temperature range. Note that while Lynden-Bell *et al*. constructed their $6H_2O/(3\times2)$

MgO(001) atomic model considering the existence of a glide plane, the genetic-algorithm-based approach led to exactly the same result without any symmetry-related constraints. Similarly, the most stable $10H_2O/(4\times2)$ MgO(001) system model was obtained without symmetry constraints, although it is characterized by the centered $c(4 \times 2)$ surface lattice. This shows the potential of the DoDo genetic algorithm code.

6. Summary

This Dissertation presents the course and the outcome of selected surface structure investigations. All these studies are examples of an outstanding cooperation between theoretical and experimental groups that led to identification of the atomic structure of the investigated model materials, *i.e.* the SiO_2/Ru(0001), the $FeSiO_4$/Ru(0001) and the H_2O/MgO(001) ultrathin films.

The Dissertation shows how computational tools give insight into the nature of experimentally obtained materials. In particular, the agreement of the properties simulated for the computationally investigated structure models with the available high-quality experimental data leads to deeper understanding of the structure-property relationship.

However, the most important scientific achievement presented in this Dissertation is the implementation of the genetic algorithm that is applicable to surface science problems. The algorithm is designed to perform global exploration of the energy landscapes of 2D-periodic systems (surfaces) and is suitable to search for atomic structure of *e.g.* reconstructed surfaces or adsorbed species.

Appendix: keywords used in the DoDo input file

A sample input file used for the DoDo genetic algorithm, `genetic_example.inp`, can be generated by calling the program with `--vasp_example` (for surfaces) or `--tmole_example` (for clusters) options. Such an input file presents all keywords implemented in the program. For clarity, these keywords are separated in sections but the file is (almost) free formatted. The hash sign # starts a comment. The keywords are followed by default values summarized in the list presented below.

General settings

population size SIZE(int)

Defines the maximum population size (*cf.* Section 4.2.3 Survival of the fittest); **SIZE** > 1; default **SIZE** = 20.

min max queued structures MIN(int) MAX(int)

Defines the minimum and maximum number of the local optimization jobs present at any moment in the queue (*cf.* Section 4.2.2 Local optimization and Section 4.2.6 Dynamic steady-state algorithm); use **MIN** = 1 to have a 'standard' generation-based GA run; **MAX** ≥ **MIN** ≥ 1; defaults **MIN** = 1, **MAX** = 10.

optimized structures MAX(int)

Defines the total number of structures optimized during the run; default **MAX** = 100.

Initial structures generator settings

gen ini max tries MAX(int)

Defines the maximum number of tries performed in order to generate a single initial structure model; default **MAX** = 1000.

initial structures NUMBER(int)

Defines the total number of generated initial structures; default **NUMBER** = 20.

atomic radii bounds A(float) B(float)

Defines the multipliers for atomic radii used to check whether two atoms are bound (*cf*. Eqns 4.2.1 and 4.2.2); defaults **A** = 0.7; **B** = 1.2

composition

SYMBOL1(string) MIN1a(int) MAX1b(int)

SYMBOL2(string) MIN2a(int) MAX2b(int)

SYMBOL3(string) MIN3a(int) MAX3b(int)

…

end composition

The composition block defines the desired atomic composition of the initial structures. **SYMBOL** is a chemical symbol of the given element. **MIN** and **MAX** define minimum and maximum number of atoms for that element (*cf*. Figure 4.3). **MAX** ≥ **MIN**.

seed file FILENAME(string)

Declares the file that contains seeds given in the internal DoDo format (*cf*. Section 4.2.1 Initial structure models); default **FILENAME** = OLD_POPULATION.dodo.

seeds amount NUMBER1(int) NUMBER2(int) NUMBER3(int) …

Defines how many initial species will be generated from each seed given in the **seed file** (Figure 4.2). The number of the values has to be the same as the count of the seeds given in the **seed file**, *i.e.* each seed in the **seed file** has to be accounted for.

penetration rate FACTOR(float)

Defines the depth of the atomic layer accessible for penetration (Figure 4.4). $1.0 \geq$ **FACTOR** ≥ 0.0; 0.0 means no penetration, 1.0 means allowing penetration to the highest frozen atom. Default **FACTOR** = 0.3.

Crossover and mutation settings

constant composition mode

This line is a sanity check. If this line is present in the input file, **MIN** should be equal to **MAX** for each **SYMBOL** in the **composition** section. Otherwise, an error is raised. Delete this line if the number of atoms is to be varied.

crossover max tries MAX(int)

Defines the maximum number of tries performed in order to generate a single structure model; default **MAX** = 1000.

fitness scaling FACTOR(float)

Defines the scaling factor α used in Eq. 4.2.3; default **FACTOR** = 1.0.

mutation probability VALUE(float)

Defines the mutation probability used only for cluster GA implementation; default **VALUE** = 0.01.

Selection settings

selection RMS VALUE(float)

Defines the maximum root mean square of the distances between the corresponding atoms for which two structures are considered similar (*cf*. Section 4.2.5 Similarity recognition); used also for cluster implementation; default **VALUE** = 0.3.

selection dstmax VALUE(float)

Defines the maximum distance [Å] used to seek for the corresponding atoms, *i.e*. the length of the boxes created during the similarity recognition procedure (Figure 4.8); used also for cluster implementation; default **VALUE** = 0.3.

Software control settings

queueing system NAME(string)

Defines which job scheduling software interface should be used. **NAME** = pbs or **NAME** = hlrn. Default **NAME** = pbs.

time interval VALUE(int)

Defines how often [seconds] the GA checks the queuing system for the presence of submitted optimization jobs. Default **VALUE** = 600.

computational platform NAME(string)

Defines which computational platform interface should be used. **NAME** = vasp or **NAME** = turbomole. In the case of **NAME** = vasp, the presence of additional DISCPAR file is recognized.

job file NAME(string)

Declares the file that is used to submit a job to the queuing system. Default **NAME** = job.run.

tmole file NAME(string)

Declares the file used by Tmole2.0 to create the control file for Turbomole. This input line is important only if **computational platform turbomole** is used. Default **tmole file =** turbo.in.

List of abbrevations

2D	Two-dimensional
3D	Three-dimensional
DoDo	Genetic algorithm package described in this dissertation
GA	Genetic Algorithm
HAGA	Hybrid *Ab initio* Genetic Algorithm for clusters
IRAS	Infrared Reflection Absorption Spectroscopy technique
LST	Linear Synchronous Transit method
MD	Molecular Dynamics method
PBS	Portable Batch System
PES	Potential Energy Surface
STM	Scanning Tunnelling Microscopy technique
QM	Quantum Mechanics (or Quantum Mechanical)
QST	Quadratic Synchronous Transit method
UHV	Ultra High Vacuum
UniCat	“Unifying Concepts in Catalysis” cluster of excellence
VASP	Vienna *Ab initio* Simulation Package
XPS	X-ray Photoelectron Spectroscopy technique

References

[1] D. Wales, *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses* (Cambridge University Press, Cambridge, 2003).

[2] I. N. Levine, *Quantum Chemistry* (Prentice Hall, 2008).

[3] H. B. Schlegel, *J. Comput. Chem.* **24**, 1514 (2003).

[4] T. A. Halgren and W. N. Lipscomb, *Chem. Phys. Lett.* **49**, 225 (1977).

[5] C. V. Ciobanu, C.-Z. Wang, and K.-M. Ho, in *Atomic Structure Prediction of Nanostructures, Clusters and Surfaces* (Wiley-VCH Verlag GmbH & Co. KGaA, 2013), p. 71.

[6] H. Jónsson, G. Mills, and K. W. Jacobsen, in *Classical and Quantum Dynamics in Condensed Phase Simulations*, edited by G. C. B. J. Berne, D. F. Coker (World Scientific, Singapore, 1998), p. 385.

[7] J. Weissenrieder, S. Kaya, J. L. Lu, H. J. Gao, S. Shaikhutdinov, H. J. Freund, M. Sierka, T. K. Todorova, and J. Sauer, *Phys. Rev. Lett.* **95**, 076103 (2005).

[8] M. Sierka, J. Döbler, J. Sauer, G. Santambrogio, M. Brümmer, L. Wöste, E. Janssens, G. Meijer, and K. R. Asmis, *Angew. Chem., Int. Ed.* **46**, 3372 (2007).

[9] P. Hünenberger, in *Advanced Computer Simulation*, edited by C. Holm and K. Kremer (Springer Berlin Heidelberg, 2005), Vol. 173, p. 105.

[10] D. M. Daven, N. Tit, J. R. Morris, and K. M. Ho, *Chem. Phys. Lett.* **256**, 195 (1996).

[11] R. G. Cook and M. A. J. Qadri, *J. Exp. Psychol.-Anim. Behav. Process.* **39**, 357 (2013).

[12] S. Algoul, M. S. Alam, M. A. Hossain, and M. A. A. Majumder, *Med. Biol. Eng. Comput.* **49**, 51 (2011).

[13] E. J. de Oliveira, G. J. Rosseti, L. W. de Oliveira, F. V. Gomes, and W. Peres, *Int. J. Elec. Power* **57**, 129 (2014).

[14] A. Fraser, *Aust. J. Biol. Sci.* **10**, 484 (1957).

[15] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Wesley Longman Publishing Co., Inc., 1989).

[16] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence* (MIT Press, 1992).

[17] K. A. D. Jong, Doctoral Dissertation, University of Michigan, 1975.

[18] H. H. Rosenbrock, *Comput. J.* **3**, 175 (1960).

[19] S. Luke, *Essentials of Metaheuristics* (Lulu, 2013).

[20] http://apps.webofknowledge.com/, access date: 5 September 2014

[21] M. J. J. Blommers, C. B. Lucasius, G. Kateman, and R. Kaptein, *Biopolymers* **32**, 45 (1992).

[22] F. Gray, Pulse code communication, United States Patent 2632058, 1953

[23] R. S. Judson, M. E. Colvin, J. C. Meza, A. Huffer, and D. Gutierrez, *Int. J. Quantum Chem.* **44**, 277 (1992).

[24] R. W. Smith, *Comput. Phys. Commun.* **71**, 134 (1992).

[25] B. Hartke, *J. Phys. Chem.* **97**, 9973 (1993).

[26] D. B. McGarrah and R. S. Judson, *J. Comput. Chem.* **14**, 1385 (1993).

[27] T. Brodmeier and E. Pretsch, *J. Comput. Chem.* **15**, 588 (1994).

[28] D. M. Deaven and K. M. Ho, *Phys. Rev. Lett.* **75**, 288 (1995).

[29] B. Hartke, M. Schütz, and H.-J. Werner, *Chem. Phys.* **239**, 561 (1998).

[30] B. Hartke, *J. Comput. Chem.* **20**, 1752 (1999).

[31] E. Curotto, A. Matro, D. L. Freeman, and J. D. Doll, *J. Chem. Phys.* **108**, 729 (1998).

[32] V. E. Bazterra, M. B. Ferraro, and J. C. Facelli, *J. Chem. Phys.* **116**, 5984 (2002).

[33] F. C. Chuang, C. V. Ciobanu, V. B. Shenoy, C. Z. Wang, and K. M. Ho, *Surf. Sci.* **573**, L375 (2004).

[34] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Science* **220**, 671 (1983).

[35] D. J. Wales and J. P. K. Doye, *J. Phys. Chem. A* **101**, 5111 (1997).

[36] D. J. Wales and H. A. Scheraga, *Science* **285**, 1368 (1999).

[37] B. Olson, I. Hashmi, K. Molloy, and A. Shehu, *Adv. Artif. Intell.* **2012**, 674832 (2012).

[38] R. Poli, *J. Artif. Evol. Appl.* **2008**, 685175 (2008).

[39] A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12562 (2002).

[40] A. Laio and F. L. Gervasio, *Rep. Prog. Phys.* **71**, 126601 (2008).

[41] G. M. Torrie and J. P. Valleau, *J. Comput. Phys.* **23**, 187 (1977).

[42] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, *J. Comput. Chem.* **13**, 1011 (1992).

[43] M. Lindroos, H. Pfnür, G. Held, and D. Menzel, *Surf. Sci.* **222**, 451 (1989).

[44] H. Pfnür, G. Held, M. Lindroos, and D. Menzel, *Surf. Sci.* **220**, 43 (1989).

[45] K. L. Kostov, M. Gsell, P. Jakob, T. Moritz, W. Widdra, and D. Menzel, *Surf. Sci.* **394**, L138 (1997).

[46] C. Stampfl, S. Schwegmann, H. Over, M. Scheffler, and G. Ertl, *Phys. Rev. Lett.* **77**, 3371 (1996).

[47] L. G. Dainyak, B. B. Zviagina, V. S. Rusakov, and V. A. Drits, *Eur. J. Mineral.* **18**, 753 (2006).

[48] G. E. Moore, *Proc. IEEE* **86**, 82 (1998).

[49] http://www.intel.com/content/www/us/en/history/museum-gordon-moore-law.html, access date: 11.09.2014

[50] http://en.wikipedia.org/wiki/FLOPS, access date: 11.09.2014

[51] M. Sierka, *Progr. Surf. Sci.* **85**, 398 (2010).

[52] B. Helmich and M. Sierka, *J. Comput. Chem.* **33**, 134 (2012).

[53] B. Hartke, *Phys. Chem. Chem. Phys.* **5**, 275 (2003).

[54] J. R. Heidberg, B.; Wetter, D., *Ber. Bunsenges. Phys. Chem.* **99**, 1333 (1995).

[55] D. Ferry, A. Glebov, V. Senz, J. Suzanne, J. P. Toennies, and H. Weiss, *J. Chem. Phys.* **105**, 1697 (1996).

[56] D. Ferry, A. Glebov, V. Senz, J. Suzanne, J. P. Toennies, and H. Weiss, *Surf. Sci.* **377–379**, 634 (1997).

[57] M. J. Stirniman, C. Huang, R. Scott Smith, S. A. Joyce, and B. D. Kay, *J. Chem. Phys.* **105**, 1295 (1996).

[58] C. Xu and D. W. Goodman, *Chem. Phys. Lett.* **265**, 341 (1997).

[59] D. Ferry, S. Picaud, P. N. M. Hoang, C. Girardet, L. Giordano, B. Demirdjian, and J. Suzanne, *Surf. Sci.* **409**, 101 (1998).

[60] L. Giordano, J. Goniakowski, and J. Suzanne, *Phys. Rev. Lett.* **81**, 1271 (1998).

[61] M. Odelius, *Phys. Rev. Lett.* **82**, 3919 (1999).

[62] J.-H. Cho, J. M. Park, and K. S. Kim, *Phys. Rev. B* **62**, 9981 (2000).

[63] L. Delle Site, A. Alavi, and R. M. Lynden-Bell, *J. Chem. Phys.* **113**, 3344 (2000).

[64] R. M. Lynden-Bell, L. Delle Site, and A. Alavi, *Surf. Sci.* **496**, L1 (2002).

[65] K. Jug, B. Heidberg, and T. Bredow, *Surf. Sci.* **601**, 1529 (2007).

[66] K. Jug, B. Heidberg, and T. Bredow, *J. Phys. Chem. C* **111**, 6846 (2007).

[67] G. Kresse and J. Furthmüller, *Phys. Rev. B* **54**, 11169 (1996).

[68] G. Kresse and J. Furthmüller, *Comput. Mater. Sci.* **6**, 15 (1996).

[69] J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).

[70] J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **78**, 1396 (1997).

[71] H. J. Monkhorst and J. D. Pack, *Phys. Rev. B* **13**, 5188 (1976).

Declaration

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Radosław Stanisław Włodarczyk

Berlin, den 26. November 2014

Complete list of the candidate's publications

(I) C. Büchner, L. Lichtenstein, X. Yu, J. A. Boscoboinik, B. Yang, W. E. Kaden, M. Heyde, S. K. Shaikhutdinov, **R. Włodarczyk**, M. Sierka, J. Sauer, and H.-J. Freund, "Ultrathin Silica Films: The Atomic Structure of Two-Dimensional Crystals and Glasses," *Chem. Eur. J.* **20**, 9176 (2014).

(II) H. Zhao, S. T. Oyama, H.-J. Freund, **R. Włodarczyk**, and M. Sierka, "Nature of Active Sites in Ni_2P Hydrotreating Catalysts as Probed by Iron Substitution," *Appl. Catal. B Environ.* **164**, 204 (2014).

(III) **R. Włodarczyk**, J. Sauer, X. Yu, J. A. Boscoboinik, B. Yang, S. Shaikhutdinov, and H.-J. Freund, "Atomic Structure of an Ultrathin Fe-Silicate Film Grown on a Metal: A Monolayer of Clay?," *J. Am. Chem. Soc.* **135**, 19222 (2013).

(IV) **R. Włodarczyk**, M. Sierka, J. Sauer, D. Löffler, J. J. Uhlrich, X. Yu, B. Yang, I. M. N. Groot, S. Shaikhutdinov, and H. J. Freund, "Tuning the Electronic Structure of Ultrathin Crystalline Silica Films on Ru(0001)," *Phys. Rev. B* **85**, 085403 (2012).

(V) L. Lichtenstein, C. Büchner, B. Yang, S. Shaikhutdinov, M. Heyde, M. Sierka, **R. Włodarczyk**, J. Sauer, and H.-J. Freund, "The Atomic Structure of a Metal-Supported Vitreous Thin Silica Film," *Angew. Chem., Int. Ed.* **51**, 404 (2012).

(VI) J. A. Boscoboinik, X. Yu, B. Yang, F. D. Fischer, **R. Włodarczyk**, M. Sierka, S. Shaikhutdinov, J. Sauer, and H.-J. Freund, "Modeling Zeolites with Metal-Supported Two-Dimensional Aluminosilicate Films," *Angew. Chem., Int. Ed.* **51**, 6005 (2012).

(VII) B. Yang, W. E. Kaden, X. Yu, J. A. Boscoboinik, Y. Martynova, L. Lichtenstein, M. Heyde, M. Sterrer, **R. Włodarczyk**, M. Sierka, J. Sauer, S. Shaikhutdinov, and H.-J. Freund, "Thin Silica Films on Ru(0001): Monolayer, Bilayer and Three-Dimensional Networks of [SiO_4] Tetrahedra," *Phys. Chem. Chem. Phys.* **14**, 11344 (2012).

(VIII) S. T. Oyama, H. Y. Zhao, H. J. Freund, K. Asakura, **R. Włodarczyk**, and M. Sierka, "Unprecedented Selectivity to the Direct Desulfurization (DDS) Pathway in a Highly Active FeNi Bimetallic Phosphide Catalyst," *J. Catal.* **285**, 1 (2012).

(IX) A. Schüller, D. Blauth, J. Seifert, M. Busch, H. Winter, K. Gärtner, **R. Włodarczyk**, J. Sauer, and M. Sierka, "Fast Atom Diffraction During Grazing Scattering from a MgO(001) Surface," *Surf. Sci.* **606**, 161 (2012).

(X) **R. Włodarczyk**, M. Sierka, K. Kwapień, J. Sauer, E. Carrasco, A. Aumer, J. F. Gomes, M. Sterrer, and H.-J. Freund, "Structures of the Ordered Water Monolayer on MgO(001)," *J. Phys. Chem. C* **115**, 6764 (2011).

(XI) A. M. Burow, T. Wende, M. Sierka, **R. Włodarczyk**, J. Sauer, P. Claes, L. Jiang, G. Meijer, P. Lievens, and K. R. Asmis, "Structures and Vibrational Spectroscopy of Partially Reduced Gas-Phase Cerium Oxide Clusters," *Phys. Chem. Chem. Phys.* **13**, 19393 (2011).

(XII) W. Unterberger, T. J. Lerotholi, E. A. Kroger, M. J. Knight, D. A. Duncan, D. Kreikemeyer-Lorenzo, K. A. Hogan, D. C. Jackson, **R. Włodarczyk**, M. Sierka, J. Sauer, and D. P. Woodruff, "Local Hydroxyl Adsorption Geometry on TiO_2(110)," *Phys. Rev. B* **84**, 115461 (2011).

(XIII) G. Mazur, M. Makowski, **R. Włodarczyk**, and Y. Aoki, "Dressed TDDFT Study of Low-Lying Electronic Excited States in Selected Linear Polyenes and Diphenylopolyenes," *Int. J. Quantum Chem.* **111**, 819 (2011).

(XIV) U. Specht, M. Busch, J. Seifert, A. Schüller, H. Winter, K. Gärtner, **R. Włodarczyk**, M. Sierka, and J. Sauer, "Rainbow Scattering Under Axial Surface Channeling from KCl(001) Surface," *Phys. Rev. B* **84**, 125440 (2011).

(XV) U. Specht, M. Busch, J. Seifert, H. Winter, K. Gärtner, **R. Włodarczyk**, M. Sierka, and J. Sauer, "Classical and Quantum Mechanical Rainbow-Scattering of Fast He Atoms from a KCl(001) Surface," *Nucl. Instrum. Meth. B* **269**, 799 (2011).

(XVI) D. Löffler, J. J. Uhlrich, M. Baron, B. Yang, X. Yu, L. Lichtenstein, L. Heinke, C. Büchner, M. Heyde, S. Shaikhutdinov, H. J. Freund, **R. Włodarczyk**, M. Sierka, and J. Sauer, "Growth and Structure of Crystalline Silica Sheet on Ru(0001)," *Phys. Rev. Lett.* **105**, 146104 (2010).

(XVII) J. Seifert, A. Schuller, H. Winter, **R. Włodarczyk**, J. Sauer, and M. Sierka, "Diffraction of Fast Atoms During Grazing Scattering from the Surface of an Ultrathin Silica Film on Mo(112)," *Phys. Rev. B* **82**, 035436 (2010).

(XVIII) G. Mazur and **R. Włodarczyk**, "Application of the Dressed Time-Dependent Density Functional Theory for the Excited States of Linear Polyenes," *J. Comput. Chem.* **30**, 811 (2009).

Curriculum Vitae

11.2008 – present **Humboldt University of Berlin**, Ph.D. studies in Theoretical Chemistry

Dissertation: *Surface Structure Predictions and Development of Global Exploration Tools.*

2003 – 2008 **Jagiellonian University, Kraków**, Poland, M.Sc. in Theoretical Chemistry.

Thesis: *Implementation of the TDDFT Method and Its Extension Aiming at Improving Energies of Excited States Having Significant Admixture of Doubly Excited Configurations.*

The Thesis earned *Professor Jacek Rychlewski Award* for the best M.Sc. Thesis in quantum chemistry in Poland (2008)

Acknowledgement

I would like to express my deepest gratitude to my two supervisors, Prof. Dr. Marek Sierka and Prof. Dr. Dr. h.c. Joachim Sauer, for their support and expert advice throughout my Ph.D. studies. In particular, I appreciate many fruitful scientific discussions that have certainly developed my critical-thinking skill.

I would like to thank all collaborators from the Chemical Physics Department of the Fritz Haber Institute of the Max Planck Society, in particular Prof. Dr. Hans-Joachim Freund and Dr. Shamil Shaikhutdinov, for the rewarding years-long cooperation.

I would like to thank Prof. Dr. Bernd Hartke, who generously agreed to review this work.

The results presented in this Dissertation would have been impossible to obtain without computers. Thus, I would like to thank our system administrator, Thomas Dargel, for his 24/7/365 service.

I would like to thank Dr. Milan Ončák for thorough and excellent feedback on earlier versions of this Dissertation.

I gratefully acknowledge financial support by the Cluster of Excellence “Unifying Concepts in Catalysis”.

Last but not least, I would like to thank my family, friends, and all co-workers who have supported me through the duration of this project. In particular, my wife Lena, whose love and encouragement were indispensable.